Year : 2014

Domain Adaptation in remote sensing: increasing the portability of land-cover classifiers

Giona Matasci

UNIL | Université de Lausanne

Faculté des Géosciences et de l'Environnement

Institut des Dynamiques de la Surface Terrestre

DOMAIN ADAPTATION IN REMOTE SENSING: INCREASING THE PORTABILITY OF LAND-COVER CLASSIFIERS

Thèse de doctorat

présentée à la

Faculté des Géosciences et de l'Environnement
de l'Université de Lausanne

par

GIONA MATASCI

B. Sc., M. Sc. Université de Lausanne, Suisse

JURY

Président du colloque:	Prof. Eric Verrecchia
Directeur de Thèse:	Prof. Mikhail Kanevski
Co-directeur de Thèse:	Dr. Devis Tuia
Expert:	Prof. Pascal Frossard
Expert:	Prof. Melba Crawford
Expert:	Prof. Paul Scheunders

Lausanne, 2014

Giona Matasci: *Domain Adaptation in remote sensing: increasing the portability of land-cover classifiers,* PhD Thesis submitted to the Faculty of Geosciences and Environment of the University of Lausanne in partial fulfillment of the requirements for the degree of Doctor in Environmental Sciences, © September 2014.

UNIL | Université de Lausanne
Décanat Géosciences et de l'Environnement
bâtiment Géopolis
CH-1015 Lausanne

IMPRIMATUR

Vu le rapport présenté par le jury d'examen, composé de

Président de la séance publique :	M. le Professeur Eric Verrecchia
Président du colloque :	M. le Professeur Eric Verrecchia
Co-Directeur de thèse :	M. le Professeur Mikhail Kanevski
Co-Directeur de thèse :	M. le Docteur Devis Tuia
Expert externe :	M. le Professeur Pascal Frossard
Experte externe :	Mme la Professeure Melba Crawford
Expert externe :	M. le Professeur Paul Scheunders

Le Doyen de la Faculté des géosciences et de l'environnement autorise l'impression de la thèse de

Monsieur Giona MATASCI

Titulaire d'une
Maîtrise universitaire ès Sciences en géosciences de l'environnement
Université de Lausanne

intitulée

DOMAIN ADAPTATION IN REMOTE SENSING : INCREASING THE PORTABILITY OF LAND-COVER CLASSIFIERS

Lausanne, le 26 septembre 2014

Pour le Doyen de la Faculté des géosciences et de l'environnement

Professeur Eric Verrecchia, Vice-doyen

DOMAIN ADAPTATION IN REMOTE SENSING: INCREASING THE PORTABILITY OF LAND-COVER CLASSIFIERS

Giona Matasci — Institute of Earth Surface Dynamics

ABSTRACT

Among the types of remote sensing acquisitions, optical images are certainly one of the most widely relied upon data sources for Earth observation. They provide detailed measurements of the electromagnetic radiation reflected or emitted by each pixel in the scene. Through a process termed supervised land-cover classification, this allows to automatically yet accurately distinguish objects at the surface of our planet. In this respect, when producing a land-cover map of the surveyed area, the availability of training examples representative of each thematic class is crucial for the success of the classification procedure.

However, in real applications, due to several constraints on the sample collection process, labeled pixels are usually scarce. When analyzing an image for which those key samples are unavailable, a viable solution consists in resorting to the ground truth data of other previously acquired images. This option is attractive but several factors such as atmospheric, ground and acquisition conditions can cause radiometric differences between the images, hindering therefore the transfer of knowledge from one image to another.

The goal of this Thesis is to supply remote sensing image analysts with suitable processing techniques to ensure a robust portability of the classification models across different images. The ultimate purpose is to map the land-cover classes over large spatial and temporal extents with minimal ground information. To overcome, or simply quantify, the observed shifts in the statistical distribution of the spectra of the materials, we study four approaches issued from the field of machine learning.

First, we propose a strategy to intelligently sample the image of interest to collect the labels only in correspondence of the most useful pixels. This iterative routine is based on a constant evaluation of the pertinence to the new image of the initial training data actually belonging to a different image.

Second, an approach to reduce the radiometric differences among the images by projecting the respective pixels in a common new data space is presented. We analyze a kernel-based feature extraction framework suited for such problems, showing that, after this relative normalization, the cross-image generalization abilities of a classifier are highly increased.

Third, we test a new data-driven measure of distance between probability distributions to assess the distortions caused by differences in the acquisition geometry affecting series of multi-angle images. Also, we gauge the portability of classification models through the sequences. In both exercises, the efficacy of classic physically- and statistically-based normalization methods is discussed.

Finally, we explore a new family of approaches based on sparse representations of the samples to reciprocally convert the data space of two images. The projection function bridging the images allows a synthesis of new pixels with more similar characteristics ultimately facilitating the land-cover mapping across images.

DOMAIN ADAPTATION IN REMOTE SENSING: INCREASING THE PORTABILITY OF LAND-COVER CLASSIFIERS

Giona Matasci — Institut des Dynamiques de la Surface Terrestre

RÉSUMÉ

Parmi les types de mesures par télédétection, les images optiques sont certainement l'une des sources de données les plus largement utilisées pour l'observation de la Terre. Elles fournissent des informations détaillées concernant le rayonnement électromagnétique réfléchi ou émis par chaque pixel de la zone étudiée. À travers un processus appelé classification supervisée, ces images permettent d'identifier de façon automatique et précise les objets à la surface de notre planète. À cet égard, lors de la production d'une carte de la couverture du sol, la disponibilité d'exemples d'entrainement représentatifs de chaque classe thématique est cruciale pour le succès de la procédure de classification.

Cependant, dans des applications concrètes, en raison de plusieurs contraintes dans la collecte des échantillons, les pixels étiquetés sont généralement rares. Lors de l'analyse d'une image pour laquelle ces exemples clés ne sont pas disponibles, une solution viable consiste à recourir aux données de terrain appartenant à d'autres images précédemment acquises. Cette option est intéressante, mais plusieurs facteurs tels que les conditions atmosphériques, au sol et d'acquisition peuvent entraîner des différences radiométriques entre les images, empêchant partiellement le transfert des connaissances d'une image à l'autre.

L'objectif de cette Thèse est de fournir aux analystes d'images de télédétection des techniques de traitement appropriées pour assurer la portabilité des modèles de classification entre les différentes images. Le but ultime est de cartographier l'occupation du sol sur de grandes étendues spatiales et temporelles à partir d'un minimum d'informations au sol. Pour corriger, ou tout simplement quantifier les changements observés dans la distribution statistique des spectres des matériaux, nous étudions quatre approches issues du champ d'études de l'apprentissage automatique.

Premièrement, nous proposons une stratégie pour échantillonner intelligemment l'image à classifier afin d'acquérir les étiquettes thématiques en correspondance que des pixels les plus utiles. Cette routine itérative est basée sur une évaluation constante de la pertinence pour la nouvelle image des données d'entrainement initiales appartenant à une image différente.

Dans un deuxième temps, nous présentons une approche pour réduire les différences radiométriques entre les images en projetant les pixels respectifs dans un nouvel espace de données commun. Des méthodes à noyaux pour la réduction de dimensionnalité adaptées pour de tels problèmes sont analysées. Il est montré qu'après cette normalisation relative, les capacités de généralisation entre images d'un classificateur sont fortement augmentées.

Ensuite, nous testons une récente mesure non-paramétrique de distance entre distributions de probabilité pour évaluer les distorsions causées par des différences dans la géométrie d'acquisition affectant des séries d'images multi-angulaires. En outre, la portabilité des modèles de classification à travers les séquences est aussi mesurée. Dans ces deux exercices, nous discutons l'efficacité des méthodes classiques de normalisation à base statistique et physique.

Enfin, nous explorons une nouvelle famille d'approches fondées sur les représentations parcimonieuses des échantillons afin de convertir réciproquement l'espace de données de deux images. La fonction de projection joignant les images permet de synthétiser de nouveaux pixels avec des caractéristiques plus proches qui faciliteront finalement la cartographie de l'occupation du sol entre des images différentes.

ACKNOWLEDGMENTS

During these four years many people contributed to this work, some directly, others indirectly, but all of them pushed me a bit forward and made this achievement possible.

Let's start with my co-advisors, Prof. Mikhail Kanevski and Dr. Devis Tuia. Mikhail has always been a great supervisor, starting from my Master thesis back in 2008 all the way through this PhD. I am very grateful to him for having introduced me to the world of research in general, and specifically to the fields of machine learning and data analysis.

The person that has played the biggest role in my leaning towards remote sensing is certainly Devis, as much a friend as an advisor. It all began with a semester project on the analysis of these "weird" images and who would have guessed this would ultimately lead to a PhD thesis! Devis, the list of things you taught me during these years is endless and I cannot find a proper way to thank you for all this.

Lots of advice for some parts of the thesis was provided by Prof. Lorenzo Bruzzone. I wish to thank him and his RSLab for hosting me for that internship and for the nice time I had in Trento (and on the surrounding ski slopes).

I am equally indebted to Fabio and Nathan at DigitalGlobe for the great collaboration we had/are having, mostly via remote video conferences, although we also had time to meet and have fun together at real conferences all around the world.

I owe a lot to Frank, a fellow PhD student at EPFL. He (and Diego, who I thank as well) showed great patience in teaching me the secrets of dictionaries and sparse representations. Also at EPFL, I am grateful to the guys of LaSIG for the frequent scientific (read "cross-lab presentations") and less scientific (read "Sat") exchanges.

I would like to acknowledge all the members of the Jury who agreed to read and evaluate this Manuscript, in particular Prof. Paul Scheunders who came to Lausanne twice to listen to me.

Furthermore, I express my gratitude to Prof. François Bavaud for introducing me to statistics and to Prof. Gustavo "Gustao" Camps-Valls for the many things I learned from him about remote sensing.

At UNIL, the colleague that deserves a special thank you is Mitch, the (lucky?) friend who had to share the office (actually three of them in two different buildings) with me for almost the entire length of the thesis. Your humor inside and outside the office will never be forgotten!

Enjoyable moments have been the norm within the GIGAR group and therefore I would like to deeply thank its past and present members, especially Loris and Mary.

A big "merci" goes to the wider IGAR family for the countless apéros, lunches, dinners, ski runs and hockey matches we did together during these years. Moreover, I cannot overlook the technical and administrative help provided by Simon, Carole and Sabrina. Thank you!

After moving to Géopolis, I appreciated the great atmosphere we have been able to build with the other members of IDyST and the corridor-mates, in particular with those endless baby-foot matches at Zélig.

Of course, a thesis cannot be accomplished without friends from outside the university. Thus, I would like to say thank you to all my friends from Ticino and in Lausanne that helped me free my mind in my spare time.

A special place in this long list is reserved for my parents, Franca and Sandro, and for my brother Nico. The constant support and encouragement I received from you during my years in Lausanne is invaluable. Also, I want to warmly thank all the relatives, particularly those in Gordemo. Grazie davvero a tutti!

Finally, the person to whom I owe the most is Livia. No combination of words can express my feelings for you. I thank you from the bottom of my heart for always having been there for me when I needed it and for making me laugh every day, even the most difficult of these past years.

Giona Matasci, September 2014

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AL	Active Learning
APEX	Airborne Prism EXperiment
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AVIRIS	Airborne Visible and Infrared Imaging Spectrometer
BRDF	Bidirectional Reflectance Distribution Function
BT	Breaking Ties
CASI	Compact Airborne Spectrographic Imager
CAVIS	Cloud, Aerosol, water Vapor, Ice, Snow
CDF	cumulative distribution function
CHRIS	Compact High-Resolution Imaging Spectrometer
CIE	International Commission on Illumination
DA	Domain Adaptation
DEM	Digital Elevation Model
DL	Dictionary Learning
DLR	Deutsches Zentrum für Luft und Raumfahrt
DN	Digital Number
EM	electromagnetic
EO-1	Earth Observing-1 Mission
ESA	European Space Agency
ETM	Enhanced Thematic Mapper
FE	Feature Extraction
FLAASH	Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes
GDA	Generalized Discriminant Analysis
GIS	Geographical Information System
GRSS	Geoscience and Remote Sensing Society
GSD	ground sample distance
HM	Histogram Matching
HSIC	Hilbert-Schmidt Independence Criterion
IR-MAD	Iteratively Reweighted-Multivariate Alteration Detection
JM	Jeffries-Matusita
KMM	Kernel Mean Matching

KPCA	Kernel Principal Component Analysis
KSC	Kennedy Space Center
LDA	Linear Discriminant Analysis
LiDAR	Light Detection And Ranging
MAD	Multivariate Alteration Detection
MISR	Multi-angle Imaging SpectroRadiometer
MMD	Maximum Mean Discrepancy
MODTRAN	MODerate resolution atmospheric TRANsmission
MSS	Multispectral Scanner
NASA	National Aeronautics and Space Administration
NIR	near-infrared
NP	non-deterministic polynomial time
OA	Overall Accuracy
OMP	Orthogonal Matching Pursuit
PA	Producer's Accuracy
PAN	panchromatic
PCA	Principal Component Analysis
PDF	probability density function
PROBA	PRoject for On-Board Autonomy
RBF	Radial Basis Function
RKHS	reproducing kernel Hilbert space
ROSIS	Reflective Optics System Imaging Spectrometer
SAR	Synthetic Aperture Radar
SNR	signal-to-noise ratio
SPOT	Satellite Pour l'Observation de la Terre
SSTCA	Semisupervised Transfer Component Analysis
SV	Support Vector
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWIR	shortwave infrared
TCA	Transfer Component Analysis
TM	Thematic Mapper
UA	User's Accuracy
VC	Vapnik-Chervonenkis
VHR	Very High Resolution
VNIR	visible and near-infrared

SYMBOLS AND NOTATION

Throughout this Thesis, scalars will appear in italic font, vectors in lowercase boldface letters and matrices in uppercase boldface letters. All vectors are considered to be column vectors.

REMOTE SENSING

λ	wavelength [nm, μm, m]	E_λ	irradiance [W·m^{-2}]
ν	frequency [Hz]	τ	transmittance $\in [0,1]$
θ	incident angle [°]	ρ	surface reflectance $\in [0,1]$
L_λ	radiance [W·sr^{-1}·m^{-2}]		

GENERAL NOTATION

$\boldsymbol{x}$	input vector (sample)	n	number of samples
y	output label	d	number of variables
X	variables (dataset)	$\mathbb{R}^d$	d-dimensional space of real numbers
Y	labels	c	number of classes
D	labeled dataset	$\mathcal{C}$	set of classes
$\mathcal{X}$	input space	cl	a given class
$\mathcal{Y}$	output space	$P(\cdot)$	probability distribution
$\boldsymbol{X}$	$n \times d$ data matrix ($d \times n$ in DL problems)	$p(\cdot)$	probability density function
$\|\| \cdot \|\|$	ℓ_2 norm	$\boldsymbol{\Sigma}$	covariance matrix
$\|\| \cdot \|\|_p$	ℓ_p norm	$\bar{\boldsymbol{x}}$	sample mean vector
$\|\| \cdot \|\|_F$	Frobenius norm of a matrix	$\boldsymbol{\mu}$	population mean vector
$\| \cdot \|$	determinant of a matrix	$\mathbf{1}$	vector of ones
$\cdot^\top$	transpose operator	$\boldsymbol{I}$	identity matrix

MACHINE LEARNING

STATISTICAL LEARNING THEORY

y^*	predicted label	$L(\cdot,\cdot)$	loss function
$f(\cdot)$	predictive function	R_{exp}	expected risk
$\mathcal{F}$	set of functions	R_{emp}	empirical risk
θ	set of generic parameters	Ω	confidence interval
Θ	parameter space	h	VC dimension

SUPPORT VECTOR MACHINES

$\boldsymbol{w}$	vector of primal variable weights	ϱ	width of the margin
b	bias of a classifier	C	cost or penalty hyper-parameter
ξ	slack variable	$\boldsymbol{\alpha}$	vector of dual sample weights

KERNELS

$K(\cdot,\cdot)$	kernel function	$\mathcal{H}$	reproducing kernel Hilbert space
$\boldsymbol{K}$	$n \times n$ kernel matrix	$\varphi(\cdot)$	RKHS mapping function
σ	Gaussian RBF and Laplace kernel width parameter		

FEATURE EXTRACTION

$\boldsymbol{u}$	primal eigenvector	$\boldsymbol{v}$	dual eigenvector
$\boldsymbol{U}$	primal eigenvector matrix	$\boldsymbol{V}$	dual eigenvector matrix
ρ	eigenvalue	m	number of retained components
$\boldsymbol{S}_{cl}$	scatter matrix of class cl	$\boldsymbol{S}_W$	within-class scatter matrix
$\boldsymbol{\Sigma}_W$	pooled within-class covariance matrix	$\boldsymbol{S}_B$	between-class scatter matrix
$\boldsymbol{H}$	$n \times n$ centering matrix		

DICTIONARY LEARNING

$\boldsymbol{D}$	$d \times K$ dictionary matrix	$\boldsymbol{c}$	vector of sparse codes
$\boldsymbol{d}$	atom of $\boldsymbol{D}$	s	sparsity level
K	number of atoms	r	reconstruction error
$\boldsymbol{C}$	$K \times n$ matrix of sparse codes		

DOMAIN ADAPTATION

GENERAL NOTATION

$\mathcal{D}$	domain	$\boldsymbol{L}$	matrix of coefficients for MMD
$\cdot_S$	subscript denoting the source domain	δ	binary selection variable
$\cdot_T$	subscript denoting the target domain		

ADAPTIVE ACTIVE LEARNING

$\boldsymbol{\omega}$	sample weights of instance weighting SVM	$\boldsymbol{v}$	sample weights during TrAdaBoost
T	joint training set $T_S \cup T_T$	ϵ	weighted training error
U	set of unlabeled candidate samples	e_i	labeling error for sample $\boldsymbol{x}_i$
t	index of TrAdaBoost iterations	β	reweighting factor
q	number of candidates to add at each AL iteration		

FEATURE EXTRACTION FOR RELATIVE NORMALIZATION

μ	TCA & SSTCA regularization parameter	$\mathcal{L}$	graph Laplacian matrix
$\boldsymbol{K}_{YY}$	kernel matrix computed on the labels	$\boldsymbol{M}$	affinity matrix
γ	SSTCA label dependence tradeoff parameter	$\mathbb{D}$	degree matrix
λ	SSTCA manifold tradeoff parameter	k	number of neighbors for $\boldsymbol{M}$
$\boldsymbol{W}$	projection matrix	$\phi(\cdot)$	common projection function
$\cdot^*$	superscript denoting elements in projected space		

CROSS-IMAGE SYNTHESIS WITH DICTIONARY LEARNING

$\cdot_x$, $\cdot_y$	subscripts denoting two generic domains $\mathcal{D}_x$ and $\mathcal{D}_y$	η	tradeoff parameter
$\mathcal{X}$, $\mathcal{Y}$	data spaces of two generic images	ζ	regularization parameter
$\boldsymbol{X}$, $\boldsymbol{Y}$	data matrices of two generic images	$\boldsymbol{a}$	final sparse codes for cross-image synthesis
$\boldsymbol{x}$, $\boldsymbol{y}$	samples (signals) of two generic images	$\boldsymbol{W}_{x\to y}$	$\mathcal{X}$ to $\mathcal{Y}$ projection matrix
		$\boldsymbol{W}_{y\to x}$	$\mathcal{Y}$ to $\mathcal{X}$ projection matrix

Part I

INTRODUCTION

1 THESIS OVERVIEW

1.1 MOTIVATION

The origins of remote sensing

The field of remote sensing provides key tools for the observation of the surface of the Earth. Ever since the launch of the Landsat 1 mission in 1972, the images provided on a regular basis by the satellites enable us to understand the many natural or anthropic phenomena impacting our environment. Well before this turning point, analogue aerial photographs and, later on, the development of digital photography have paved the way for such a rapidly growing and fascinating discipline. Remotely sensed images constitute one of the most important sources of information to describe the spatial distribution of the objects and land-covers at the Earth's surface. The historical archives of images that are continuously complemented with the new acquisitions allow a constant assessment of the evolution of the landscape, with several valuable and diverse real-life applications. In fact, frequently updated, spatially extensive though precise imagery is paramount in assisting the development of cartographic products, conceiving land management systems, monitoring different types of natural hazards, to name but a few examples.

A wealth of data

At first, satellites were carrying sensors bearing a low to moderate spatial resolution, capturing synoptic views of large areas of the globe. The last two decades, instead, have seen the launch of a variety of satellites with on board high to very high resolution sensors providing images with an unprecedented spatial detail. Moreover, the amount of sensed data has grown extraordinarily also because of the increase in the number of spectral channels used by the sensors both in satellite and airborne missions to finely sample the electromagnetic spectrum.

Much needed analysis tools

Nowadays, such an evolution requires the proper tools to efficiently treat this large quantity of data. In this respect, the latest advances in the fields of signal and image processing have proven success in answering this demand. The early approaches to information extraction from remotely sensed images based on a visual interpretation by the analyst are more and more replaced by computer-based procedures. The key to success for such automatic routines is their ability to cope with issues like the high dimensionality of the pixels, the presence of noise in the images or the scarcity of available reference information to calibrate the models.

Ground truth collection issues

This last aspect is particularly delicate when considering the classification of the images to identify the types of land-cover appearing in the scene. Indeed, the collection of ground truth data is among the crucial factors influencing the quality of a land-cover map. To be effective, a supervised classifier needs examples suitably representing the spectral signature of the thematic classes found in the image. In an ideal situation, this set of ground truth samples is acquired for every image the user intends to analyze. Nevertheless, the sampling process is not a trivial task. The procedure, depending upon the type of application and the type of image, requires either expensive terrain campaigns or time-consuming photo-interpretation analyses. The former solution is adopted especially when dealing with hyperspectral data of low spatial resolution, whereas the latter concerns in particular studies with very high resolution images. In some critical cases, gathering new field data is simply impossible, for instance because of an inaccessible area (e. g. dangerous or politically unstable regions) or due to time constraints when a quick map update is required (e. g. emergency situations after natural disasters).

Adaptation to re-use existing ground truth

In this context, the aforementioned archives of already existing acquisitions could be conveniently exploited to alleviate the demand for reference data. If a ground truth collection has been carried out for a previous study with similar objectives, this labeled data could be profitably re-utilized, provided that the images involved share some key characteristics. The sensors having acquired them should cover the same region of the spectrum, while the imaged geographical areas should, of course, display the same type of landscape. When these conditions are met, it is likely that a classification model appropriately mapping the land-cover on the first image could perform satisfactorily on the second as well. Yet, the images have likely been acquired at different time instants and/or at distant spatial locations. The radiometric differences among them can be passably large, owing to seasonal effects, changes in the atmospheric conditions or different acquisition geometries. Therefore, an even more attractive and elegant solution consists in devising adaptation strategies to ultimately ensure that the model suitably adjusts to the new image one is interested in.

1.2 OBJECTIVES

Machine learning & remote sensing

The purpose of this Thesis is to find effective solutions to the adaptation problem outlined above. We aim at providing dedicated processing techniques to enable image analysts to easily map the land-cover over large spatial and temporal extents by leveraging already existing models or ground truth data. To this end, we take advantage of the latest developments in the field of machine learning, a discipline born at the confluence of statistics and computer science. In particular, the methodological framework in which this work sits is that of domain adaption, the branch devoted to the study of strategies to overcome a change in the statistical distribution of the datasets. As regards machine learning, methods for classification, regression and clustering have been widely utilized to analyze single remotely sensed images, proving to be decisive in applications of land-cover mapping, biophysical parameter estimation, etc. However, the fundamental problem of adaptation between several images has often been left aside.

Benefits for both disciplines

We believe that joining the efforts of the more method-oriented field of domain adaptation and those of the more application-oriented field of remote sensing can be highly beneficial for both communities. The challenges to face are basically the same, just the perspective is slightly different. Machine learning researchers studying adaptation strategies are introduced to a new stimulating field of application and will receive useful feedbacks regarding the concrete issues to focus on. On the contrary, remote sensing scientists are supplied with innovative advanced methods to solve the problems they always faced relying on techniques whose limitations were sometimes apparent.

Two solutions for adaptation

The image analysis needs in Earth observation are often dictated by the lack of or by the difficulty in obtaining ground truth data. In this respect, two possible ways to handle a cross-image thematic classification exercise exist. The practitioner either decides to spend some resources to sample each new image he receives or prefers to only rely on the already available reference data belonging to other acquisitions. These two approaches correspond to two families of adaptation methodologies developed in machine learning: supervised or unsupervised domain adaptation approaches. In this dissertation, we will analyze and discuss the advantages and shortcomings of both solutions.

Supervised domain adaptation

On the one hand, supervised adaptation approaches allow to take full advantage of the opportunity to access additional samples in the new acquisitions. In this case, the end-user is supposed to have a budget to collect a limited number of samples in the analyzed scene. As some known spectral signatures of the new image at hand are being exploited, supervised adaptation procedures are supposed to yield very accurate thematic classification products.

Unsupervised domain adaptation

On the other hand, unsupervised adaptation methodologies do not require any thematic information associated with the new images. Such a freedom opens additional opportunities if compared to the previous approach. If the appropriate normalization techniques are available, the user could successfully apply on the new images an already trained model from a previous collection. However, as no new samples can be used to refine the model, this comes to the detriment of the accuracy of the thematic maps. The field of domain adaptation proves useful to answer these needs since the change of the feature representation of the datasets is a topic of great interest.

Dataset shift assessment

A central aspect contributing to the design of both the mentioned types of adaptation approaches is represented by a correct and thorough understanding of the processes causing the change in data distributions. Thus, a further objective of this Thesis is to provide robust tools to detect and assess the dataset shift. Again, we will take advantage of the recent developments in the field of machine learning to yield a data-driven solution to be concretely used by remote sensing experts when interpreting their data.

Angular effects & compensation

Additionally, it is worth noting that among the factors affecting the radiometric homogeneity of the images, the geometry of a remotely sensed acquisition is one of the most crucial. This is particularly evident when working with high spatial resolution images. Different combinations of satellite and sun positions with respect to the scene control several fundamental physical phenomena. Therefore, another goal of this research is to analyze the angular effects ensuing from images having been acquired with different off-nadir view angles and to study the best strategies to overcome the problem. With this purpose in mind, we deem essential to evaluate the compensation ability of physical and statistical approaches traditionally used in Earth observation. We also discuss the synergy of the two approaches in identifying the reasons behind the observed spectral drift.

Summary of the objectives

To sum up, the main objectives of this Thesis can be stated as follows.

1. Increase the portability of land-cover classifiers by investigating both:
   - supervised domain adaptation strategies and
   - unsupervised domain adaptation strategies.
2. Evaluate the dataset shift affecting remote sensing images with suitable statistical measures.
3. Study and overcome the angular effects related to the geometry of the acquisition with a multidisciplinary approach based on machine learning/statistics and physics.

Large-scale applications

The application of the approaches investigated in this Thesis is twofold. First, the joint analysis of series of spatially and/or temporally spaced images can be improved. Second, we also address the related issue arising from the usage of training data sampled in small localized regions of large images. An adjustment is thus required to increase the generalization power of the classification rules learned on these shifted datasets. Eventually, a more accurate and automated large-scale mapping will be within reach in both cases, with practical applications including, for instance, the study of global urbanization trends or the continuous monitoring and timely assessment of natural hazards.

1.3 CONTRIBUTIONS OF THE THESIS

The key contributions of this Thesis will be briefly outlined in this Section. These contributions are all directly linked to the objectives discussed in the previous Section. They will be presented in four separate Chapters in Part iii of this manuscript. In addition, the relevant publications prepared during this Thesis project that are connected with each topic will be listed after each description.

1.3.1 *Chapter 6: SVM-based adaptive Active Learning via sample reweighting*

The first contribution is represented by the study of a supervised domain adaptation strategy to smartly sample the newly acquired images. In this Chapter, we propose a scheme to optimally direct the collection of new ground truth pixels based on an initial training set available from a different yet related image. The adaptive procedure separately reweights the samples of the two images based on their representativeness of the land-cover classes in the new image. At the same time it suggests which pixels are the most useful to label in order to achieve a maximal improvement of the current classification model for an effective land-cover mapping. This type of routine provides the end-user with a list of priorities allowing to minimize the additional sampling efforts.

The findings of this Chapter have been published in:

> [Matasci et al., 2012] G. Matasci, D. Tuia, and M. Kanevski. SVM-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):1335–1343, 2012.

The following work is also related to this study:

> [Matasci et al., 2011a] G. Matasci, D. Tuia, and M. Kanevski. Domain separation for efficient adaptive active learning. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3716–3719, Vancouver, Canada, 2011a.

1.3.2 *Chapter 7: Kernel-based Feature Extraction for relative radiometric normalization*

The second topic addressed during the Thesis concerns an unsupervised adaptation approach. This Chapter investigates strategies to statistically align the images in a common subspace constructed anew so that radiometrically shifted acquisitions can find a correspondence. We analyze a feature extraction technique aiming at reducing the distance between the probability distributions of the images. This type of transformation can be thought of as a relative image-to-image normalization approach. After the projection, the model portability among acquisitions is thus facilitated. As no labels from the targeted images are required, practitioners are enabled to rapidly apply on new imagery a thematic classifier they have trained beforehand.

This Chapter is based on the following accepted paper:

> [Matasci et al., Accepted.] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia. Semisupervised Transfer Component Analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, Accepted.

The following works are also related to this study:

> [Matasci et al., 2011b] G. Matasci, M. Volpi, D. Tuia, and M. Kanevski. Transfer Component Analysis for domain adaptation in image classification. In *Proceeding of the SPIE Remote Sensing conference on Image and Signal Processing for Remote Sensing*, Prague, Czech Republic, 2011b.
>
> [Volpi et al., 2012a] M. Volpi, G. Matasci, D. Tuia, and M. Kanevski. Enhanced change detection using nonlinear feature extraction. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6757–6760, Munich, Germany, 2012a.
>
> [Matasci et al., 2013a] G. Matasci, L. Bruzzone, M. Volpi, D. Tuia, and M. Kanevski. Investigating feature extraction for domain adaptation in remote sensing image classification. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Barcelona, Spain, 2013a.
>
> [Volpi et al., In press.] M. Volpi, G. Matasci, M. Kanevski, and D. Tuia. Semi-supervised multiview embedding for hyperspectral data classification. *Neurocomputing*, In press.

1.3.3 *Chapter 8: Assessing angular dataset shift and model portability in multi-angle image sequences*

This Chapter is devoted to the study of the angular properties of remote sensing image acquisitions. We aim at detecting the many physical phenomena that cause distortions in the imagery when the acquisitions take place with skewed geometries. In order to isolate the impact of the effects related to the view angle, we resort to multi-angle sequences quasi-simultaneously acquired by the satellite. We quantify the dataset shift by means of a nonlinear measure of distance between probability distributions. Furthermore, adopting an unsupervised domain adaptation setting, we assess the ability to port across the entire sequence a classification model developed on one single specific image. In this context, we shed light on the suitability of standard both absolute and relative normalization methods to overcome the observed angular shift and we analyze their combined use.

This Chapter is based on a submitted paper that is now under review:

> [Matasci et al., Submitted.] G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia. Understanding angular effects in VHR in-track multi-angle image sequences and their consequences on urban land-cover model portability. *ISPRS Journal of Photogrammetry and Remote Sensing*, Submitted.

The following work is also related to this study:

> [Matasci et al., 2013b] G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia. Statistical assessment of dataset shift and model portability in multi-angle in-track image acquisitions. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4134–4137, Melbourne, Australia, 2013b.

1.3.4 *Chapter 9: Cross-image synthesis with dictionaries*

This Chapter presents the last and most recent contribution of this Thesis. We focus again on the fundamental problem of changing the data space of the images to make them more similar to each other. This time, we propose to apply an algorithm that takes advantage of sparse representations of the images. Set in a supervised domain adaptation context, the methodology seeks a mapping function directly linking the two representations that ultimately permits to re-synthesize the pixels of a given image as though they were generated under the conditions found on another image.

This Chapter will appear in:

> [Matasci et al., 2014] G. Matasci, F. de Morsier, M. Kanevski, and D. Tuia. Domain adaptation in remote sensing through cross-image synthesis with dictionaries. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Québec City, Canada, 2014.

The following work is also related to this study:

> [Marcos Gonzalez et al., 2014] D. Marcos Gonzalez, F. de Morsier, G. Matasci, D. Tuia, and J.-P. Thiran. Hierarchical sparse representation for dictionary-based classification of hyperspectral images. In *Proceedings of the IEEE Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS)*, Lausanne, Switzerland, 2014.

1.4 ORGANIZATION OF THE MANUSCRIPT

This Thesis manuscript comprises four distinct parts structured as follows. After this introductory Chapter, the remainder of Part i consists of Chapter 2 presenting the discipline of remote sensing and the related challenges that are being faced nowadays. In the more theoretical Part ii, Chapter 3 introduces machine learning and the associated techniques that will be utilized as starting point for the experimental part of this Thesis. Chapter 4 follows with a thorough overview of the sub-field of domain adaptation. Next is the core of the dissertation: Part iii, featuring the adaptation approaches we propose. Chapter 5 firstly sets the context for adaptation studies when applied to remote sensing and reviews the state-of-the-art of the current research. The four subsequent Chapters 6, 7, 8, 9 report the respective findings previously outlined in Section 1.3. Finally, Part iv, with its conclusive Chapter 10, summarizes the main achievements and discusses the possible future research directions in the field.

2 REMOTE SENSING AND EARTH OBSERVATION

Outline: *This second Chapter will delineate the fundamental principles of the remote sensing technology as well as the main challenges arising from its latest developments. In Section 2.1, we introduce the types of sensors that can be mounted on remote platforms and discuss the possible real-life applications of this relatively new discipline. Subsequently, Section 2.2 outlines the physics and the related concepts describing the transfer of the electromagnetic radiation occurring during a remote sensing acquisition. In Section 2.3, passive remote sensing systems will be described in more detail and the principal satellite/airborne missions will be reviewed. Next, with Section 2.4 we will look at the products ultimately issued from remotely sensed images imagery, with special attention to land-cover classification. A concise literature review of the latest developments in the field will also be provided. Section 2.5 discusses the opportunities offered by jointly using multiple images, details the related radiometric issues and describes the most common answers to such problems.*

2.1 INTRODUCTION

Definition

Generally speaking, *remote sensing* can be defined as the act of measuring (sensing) the properties of objects at the surface of the Earth by means of a data collection platform not in direct contact with them (remote). The nature of the sensed signals ultimately allowing to derive the mentioned object properties can be multifaceted: optical, microwave, acoustical, etc. [Schowengerdt, 2007].

Active vs. passive sensors

The systems designed to acquire such signals can be divided into two categories, depending upon the type of interaction with the target. On the one hand, there are *passive* remote sensing instruments collecting the solar *electromagnetic* (EM) radiation that is reflected or spontaneously emitted by the Earth's surface. Among such devices we find *multispectral* and *hyperspectral* sensors [Richards and Jia, 1999] that record the energy in the visible, infrared and thermal range of the EM spectrum. On the other hand, *active* remote sensing instruments possess an artificial source of radiation, an antenna, that sends EM signals towards the Earth. The radiation that is scattered back by the objects on the ground is then detected by the sensor. In this category we find radar systems working in the microwave domain such as *Synthetic Aperture Radar* (SAR) [Curlander and McDonough, 1991] or technologies such as *Light Detection And Ranging* (LiDAR) aimed at illuminating the objects with laser beams [Shan and Toth, 2008, Wehr and Lohr, 1999]. This Thesis focuses on the processing of optical images (visible through thermal wavelengths). Thus, in the following Sections 2.2 and 2.3, we restrict the introduction of the physics of the image acquisition to this type of signals and the description of the imaging systems to passive instruments.

Georeferenced image

The above measurements are recorded by a detector arranging the collected signals in a set of cells called *pixels* forming a rectangular regular grid, the *image*. The final product output by the remote sensing image acquisition chain is a spatially georeferenced image that can be integrated, for instance, in a *Geographical Information System* (GIS) for further analyses involving additional spatial layers. Generally, the sensors are mounted on an aircraft or on a satellite collecting thus the energy *at-a-distance* instead of *in-situ*. Especially in the case of spaceborne sensors, this enables the monitoring of vast portions of the surface of our planet within short time-frames.

Applications

For all these reasons, the applications of remote sensing for Earth observation are numerous and can be pursued at a large scale [Schowengerdt, 2007]:

- monitoring of natural hazards (earthquakes, landslides, etc.) [Mantovani et al., 1996, Joyce et al., 2009, Jaboyedoff et al., 2012, Hilley et al., 2004]

- urban studies (land-use/land-cover mapping, urban growth assessment, etc.) [Jensen and Cowen, 1999, Weng, 2012, Manakos and Braun, 2014]

- agriculture (mapping of crop types and crop condition, yield predictions, etc.) [Moran et al., 1997, Lamb and Brown, 2001]

- ecological and environmental assessment (biodiversity, hazardous waste disposal, etc.) [Asner, 1998, Nagendra, 2001, Well et al., 1994]

- change detection (deforestation, melting glaciers, etc.) [Mas, 1999, Achard et al., 2002, Raup et al., 2007]

- resources exploration and monitoring (minerals, oil, natural gas, etc.) [Sabins, 1999, Brekke and Solberg, 2005]

- meteorology (climate change, weather prediction, atmosphere composition, etc.) [Yang et al., 2013, Kidder and Vonder Haar, 1995]

- mapping (regional land-cover mapping, extraction of topographic information, etc.) [Manakos and Braun, 2014, Rabus et al., 2003].

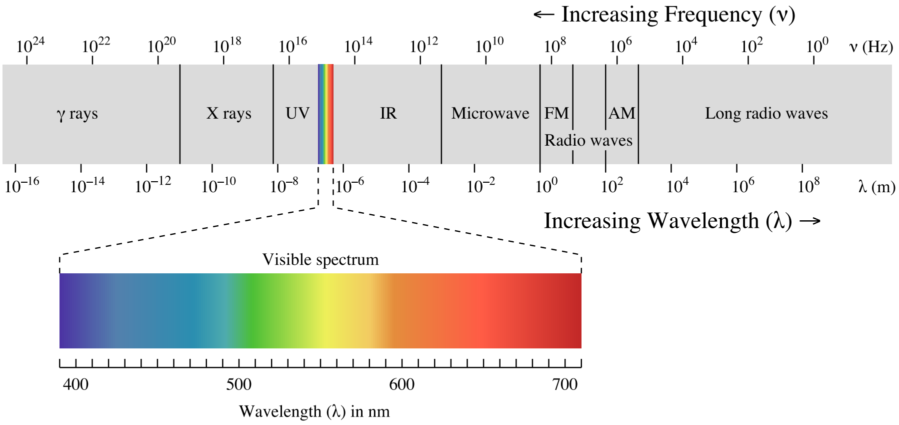


Figure 2.1: The EM spectrum. Adapted from `http://commons.wikimedia.org/`.

2.2 RADIATIVE TRANSFER

2.2.1 *What is measured by the sensor?*

EM spectrum

The solar or thermal radiation reflected or emitted by the materials constituting the surface of the Earth can be separated into categories following the regions of EM spectrum: visible (0.4 — 0.7 μm), near-infrared (NIR) (0.7 — 1.1 μm), shortwave infrared (SWIR) (1.1 — 2.5 μm), midwave infrared (3 — 5 μm), thermal or longwave infrared (8 — 14 μm). A diagram depicting the main regions of the EM spectrum with wavelengths and corresponding frequency values is reported in Fig. 2.1.

At-sensor radiance

In the *visible and near-infrared* (VNIR) to SWIR region, the part of the spectrum we will mainly be dealing with in this manuscript, the radiation transfer occurring during a remote sensing image acquisition is controlled by three distinct components, all of which depend on the wavelength λ [Schowengerdt, 2007]:

- the unscattered surface-reflected radiation, L_λ^{su}
- the down-scattered surface-reflected skylight, L_λ^{sd}
- the up-scattered path radiance, L_λ^{sp}.

Making use of these terms, the total *at-sensor radiance* L_λ^{s} reaching the platform is simply defined as

$$L_\lambda^{s} = L_\lambda^{su} + L_\lambda^{sd} + L_\lambda^{sp} \; . \tag{2.1}$$

This quantity is what is eventually measured by the sensor in [W·sr^{-1}·m^{-2}] units. Its counterpart in the thermal region of the spectrum, the total at-sensor radiance from surface emissions, will not be considered here. Figure 2.2 graphically illustrates the components of the total at-sensor radiance as well as two other physical quantities described in the paragraphs below.

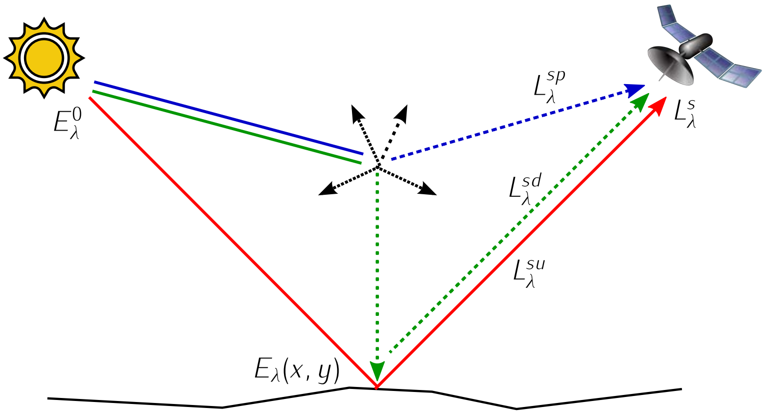


Figure 2.2: Illustration of the main physical quantities controlling the radiative transfer of a remote sensing acquisition.
red ⟶: unscattered surface-reflected radiation L^{su}_λ.
green - - →: down-scattered surface-reflected skylight L^{sd}_λ.
blue - - →: up-scattered path radiance L^{sp}_λ.

2.2.2 *Components of the at-sensor radiance*

Incident irradiance

The first term of (2.1), the unscattered surface-reflected radiation L^{su}_λ, first depends on the initial energy coming from the sun that reaches the top of the atmosphere, i. e. the spectral irradiance E^0_λ. The irradiance actually attaining the Earth's surface, denoted with E_λ, is controlled by the solar path (from the sun to the ground) atmospheric transmittance $\tau_s(\lambda)$. The latter represents the fraction of the irradiance E^0_λ that is able to make its way through the atmosphere. At this time, another factor influencing the process is the incidence angle with which the solar radiation reaches the surface of the globe in a given point (x, y) of it. For instance, a maximum reflection occurs in case of a surface perpendicular to the incoming EM beam. Therefore, the final incident irradiance is computed as follows:

$$E_\lambda(x, y) = \tau_s(\lambda) E^0_\lambda \cos(\theta(x, y)) \ , \tag{2.2}$$

where $\theta(x, y)$ is the incident angle, i. e. the angle between the solar vector and the surface normal vector at the coordinates (x, y).

Surface reflectance & surface radiance

At the surface level, the actual *surface radiance* L_λ scattered back toward the sensor is the result of a rescaling of the incident E_λ by a factor translating a crucial property of the materials, the *surface reflectance* $\rho(\lambda)$, a unitless quantity between 0 and 1:

$$L_\lambda = E_\lambda \frac{\rho(\lambda)}{\pi} \ . \tag{2.3}$$

This is valid for surfaces with an equal reflection in all the directions of the hemisphere (Lambertian surfaces). In most real-life situations this is rarely the case, since anisotropic behaviors are observed for many types of surfaces. Hence, the $\rho(\lambda)/\pi$ values are substituted by a *Bidirectional Reflectance Distribution Function* (BRDF) [Simmer and Gerstl, 1985, Schott, 2007, Schaepman-Strub et al., 2006]. Such a function describes the ratio of outgoing to incoming radiance as a function of incident (related to the sun) and view (related to the sensor) angles.

Unscattered surface-reflected radiation

Finally, since the outgoing radiance still has to traverse the atmosphere on its way back to the sensor, to obtain the unscattered surface-reflected at-sensor radiance, we apply a rescaling by the view path (from the ground to the sensing platform) transmittance $\tau_v(\lambda)$:

$$L_\lambda^{su} = \tau_v(\lambda) L_\lambda \ . \tag{2.4}$$

The view path transmittance depends upon the geometry of the acquisition, as lower values are expected for high off-nadir view angles (low satellite elevation angles) with respect to nadir acquisitions. Indeed, the optical path through the atmosphere is longer in the first situation than in the second. Moreover, note that both transmittance quantities $\tau_s(\lambda)$ and $\tau_v(\lambda)$ (proportions between 0 and 1) are highly dependent on the wavelength, as wide radiation absorption bands (around 1.4 and 1.9 μm) occur in the region of the spectrum we are considering. The presence of water vapor and carbon dioxide in the atmosphere are among the main reasons behind such a phenomenon.

Surface-reflected skylight & up-scattered path radiance

The second term of (2.1), the down-scattered surface-reflected skylight L_λ^{sd}, accounts for the radiance scattered towards the object by the atmosphere and then reflected upward. The fact that shadowed areas do not appear as completely black is an evidence of this diffuse down-scattering. Instead, the third term of (2.1), the up-scattered path radiance L_λ^{sp}, relates to the radiation directly reflected at the sensor by the atmospheric layer. For image collections taking place with slanted geometries, that is with important off-nadir view angles (long optical path), the path radiance will heavily impact the total radiance L_λ^s measured by the detectors. Both L_λ^{sd} and L_λ^{sp} are governed by molecular small-scale *Rayleigh scattering*, and by aerosol and particulate (smog, haze, etc.) *Mie scattering*. The former type of scattering more than the latter strongly depends on the wavelength, with shortwave radiations (blue and ultraviolet light) being the most affected.

2.3 PASSIVE REMOTE SENSING IMAGING SYSTEMS

2.3.1 *The hypercube*

From radiances to DNs

A passive sensor acquires a remote sensing image by sampling the continuous EM spectrum within specific wavelengths and recording the associated radiance values. This results in images having multiple spectral bands (or channels), each one responsible for different adjacent parts of the spectrum. Sensors possessing a limited number of wide bands produce multispectral images, whereas instruments with many (more than one hundred) very narrow bands yield hyperspectral images. Moreover, sensors bearing a single spectral channel having a bandwidth covering the entire VNIR range acquire *panchromatic* (PAN) images. The *spectral resolution* of a remote sensing

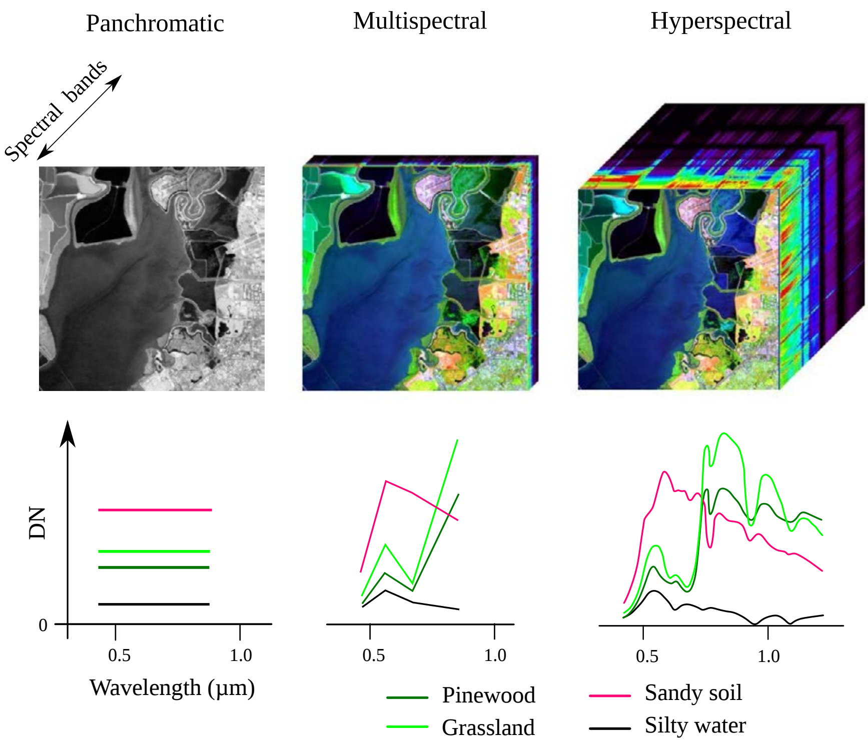


Figure 2.3: Hypercubes and associated spectral signatures produced by panchromatic, multispectral and hyperspectral sensors. Top row shows the images with a different number of spectral bands. Bottom row shows examples of the spectral signatures of four land-cover classes. Adapted from http://commons.wikimedia.org/.

sensor is defined as the bandwidth of the featured channels (in [nm] generally). This type of data can be thought of as a three-dimensional hypercube consisting of two spatial coordinates with the spectral wavelength as the third dimension. The actual physical quantity measured by the instrument, the at-sensor radiance L^s_λ, is converted and stored as an integer value, the *Digital Number* (DN), usually coded in 8, 11 or 12 bits formats (to obtain the desired *radiometric resolution*).

Spectral signature

Each pixel of the image is described by a sequence of its DN values in the different bands, the *spectral signature*. Depending on the ground-cover constituting the pixel, the spectral signatures can be very different, allowing to finely discriminate the materials. This is exactly the key opportunity offered by remote sensing systems that is shared with the field of *spectroscopy* [Clark and Roush, 1984]: the user is enabled to accurately recognize the materials or to derive meaningful parameters describing them.

An example

Figure 2.3 shows an illustration of the hypercubes resulting from acquisitions with the above-mentioned types of sensors. We present the related spectral signature for pixels belonging to the following four land-cover classes: "pinewood", "grassland", "sandy soil", "silty water". A panchromatic sensor (1st column of the scheme) outputs a gray-scale image composed of a single band in which each pixel is described by one DN value representing the average radiance recorded over a large portion of the spectrum. A multispectral system (2nd column) instead images the scene through multiple spectral channels. When the VNIR region of the spectrum is involved a true color RGB composition of the bands can be created. Each land-cover is described more precisely with several DN values. Hyperspectral sensors (3rd column) collect a large number of bands yielding a hypercube with many layers along the spectral dimension. The measured signatures are very detailed, approaching thus the true reflectance profiles of the considered materials. This permits for instance to effectively discriminate thematic classes such as "pinewood" and "grassland" that were spectrally very similar on both the panchromatic and multispectral images.

2.3.2 *Types of sensors*

Scanning systems and spatial resolution

The imaging systems mounted on airborne or spaceborne platforms can be divided into two categories with respect to the scanning procedure utilized to sense the scene. The acquisition always takes place with an *in-track* motion of the platform (along the flight path). On one side we have *pushbroom* scanners (e. g. SPOT, QuickBird) that image the full *swath width* with a linear array of detector elements. On the other side, *whiskbroom* systems (e. g. Landsat) achieve a sequential *cross-track* (perpendicular to the flight line) scan by rotating the series of detectors aligned in-track [Schowengerdt, 2007]. In the resulting image, for both types of systems, the pixel centers are spaced by the *ground sample distance* (GSD), a property often referred to as the *spatial resolution* of the image.

VHR multispectral sensors

In this respect, we distinguish the recently launched *Very High Resolution* (VHR) sensors (e. g. QuickBird, WorldView-2), producing images with a metric or sub-metric spatial resolution (< 5 m), from the previous moderate resolution sensors (e. g. MSS/TM/ETM on board the Landsat satellites, ASTER on board the Terra satellite) bearing a decametric spatial resolution [Richards and Jia, 1999]. VHR sensors yield images with a high spatial detail, enabling the end-user with the resources to carry out case studies inconceivable until the end of the 1990s. However, such high spatial resolution is obtained through a compromise: VHR devices are normally multispectral instruments with bands that are rather broad and mainly cover the VNIR region of the EM spectrum. This type of sensor is often mounted with a panchromatic instrument allowing to achieve the smallest GSD. The PAN band can then be exploited to enhance the spatial resolution of the rest of the bands by means of *pansharpening* techniques [Brower and Laben, 2000]. Another asset of the advent of VHR sensors is related to the higher *temporal resolution* of the image collections. Indeed, the increased agility of the platforms produces shorter (< 3 days) satellite revisit times (temporal interval between two consecutive acquisitions of the same scene). Finally, note that the recent launch of the WorldView-3 satellite marks a significant change in the spectral properties of VHR instruments. With its 8 SWIR channels at a 3.7 m resolution complementing the 8 VNIR bands already used by

Table 2.1: VHR satellites and their sensor characteristics. All native GSD and swath width figures refer to nadir acquisitions. Band names: C = coastal, B = blue, G = green, Y = yellow, R = red, RE = red edge, NIR = near-infrared, NIR2 = near-infrared 2. WorldView-3 CAVIS channels are dedicated calibration bands. For non-US government customers the imagery must be resampled so that the output GSD is $\geq$ 0.5 m. Adapted from `http://eijournal.com/2012/buying-optical-satellite-imagery` and `http://www.satimagingcorp.com/satellite-sensors.html`.

Satellite	Launch year	Swath width [km]	Native PAN GSD [m]	Output PAN-VNIR(-SWIR) GSD [m]	Bands
IKONOS	1999	11.3	0.82	1 - 4	PAN + 4 VNIR (B, G, R, NIR)
QuickBird	2001	16.5	0.61	0.6 - 2.4	PAN + 4 VNIR (B, G, R, NIR)
SPOT-5	2002	60	5	2.5 - 10 - 20	PAN + 3 VNIR (G, R, NIR) + 1 SWIR
WorldView-1	2007	17.6	0.5	0.5	PAN
GeoEye-1	2008	15.2	0.41	0.5 - 2	PAN + 4 VNIR (B, G, R, NIR)
WorldView-2	2009	16.4	0.46	0.5 - 2	PAN + 8 VNIR (C, B, G, Y, R, RE, NIR, NIR2)
Pléiades 1	2011	20	0.70	0.5 - 2	PAN + 4 VNIR (B, G, R, NIR)
SPOT-6	2012	60	1.5	1.5 - 6	PAN + 4 VNIR (B, G, R, NIR)
WorldView-3	expected in 2014	13	0.31	0.5 - 1.2 - 3.7	PAN + 8 VNIR + 8 SWIR (+ 12 CAVIS)

WorldView-2, this satellite can be deemed the first spaceborne platform carrying a *superspectral* VHR sensor. A list of the main VHR satellites/sensors with their characteristics is reported in Tab. 2.1.

Hyperspectral sensors

Considering the spectral resolution of passive sensors, at the opposite end, we find hyperspectral sensors capable to acquire up to hundreds of narrow bands finely sampling the spectrum [Goetz et al., 1985, Plaza et al., 2009]. Such type of instruments, when mounted on satellite platforms, never reaches the spatial resolution of multispectral sensors. In fact, since the spectral resolution is much higher ($\approx$ 10 nm wide bands), to register enough energy for each pixel, the GSD needs to be larger ($>$ 15-20 m). The list of spaceborne hyperspectral sensors is quite short. The two main instances consist of Hyperion (30 m GSD, 220 spectral bands in the VNIR to SWIR region) [Folkman et al., 2001] on board NASA's EO-1 satellite and CHRIS (highest GSD of 17 m, programmable up to 62 spectral bands in the VNIR region) [Barnsley et al., 2004] on board ESA's PROBA satellite. On the contrary, for budgetary reasons, the development of airborne sensors has been much more dynamic. In this category, a non-exhaustive list of instruments comprises ROSIS (115 VNIR spectral channels) [Mueller et al., 2002] operated by DLR, AVIRIS (224 VNIR to SWIR spectral channels) [Vane et al., 1993] developed by NASA, APEX (up to 534 VNIR to SWIR bands) [Itten et al., 2008] manufactured by ESA, as well as HyMap (126 VNIR to SWIR spectral channels) [Cocks et al., 1998] and CASI (up to 288 VNIR spectral channels) [Babey and Anger, 1993] engineered by private companies. Depending on the flight height of the airplane, the spatial resolution of the acquisitions can be very high (< 1 m).

Multi-angle sensors

A special class of imaging systems concerns platforms possessing multi-angular capabilities. Indeed, certain instruments have been designed with a collection mode allowing the acquisition of a sequence of images of the same scene with different view angles, an ability that has proved beneficial in many applications (see Chapter 8). This allows to study the BRDF properties of the materials to better understand their nature or to discriminate them more accurately. Among the spaceborne platforms offering such a flexible monitoring system there is WorldView-2. Its in-track collection mode builds on the MISR [Diner et al., 1998] and the CHRIS missions, carrying multispectral and hyperspectral sensors, respectively. As the recently designed on-board control systems have enabled a rapid re-targeting of the sensor to a wide range of look angles, the angular sampling density has been highly augmented with respect to earlier systems. Within a time frame of a few minutes, WorldView-2 can acquire angular sequences of tens of images of the same scene along an in-track collection path with off-nadir angles up to 45°.

Some examples of the images acquired by the instruments mentioned in this Section can be found in Appendix B reporting the datasets used in this Thesis.

2.4 THEMATIC IMAGE CLASSIFICATION

2.4.1 *Interpretation of the scene*

From photo-interpretation to image processing

The final goal of the processing of remote sensing imagery is the extraction of meaningful information helping the end-user in understanding the natural or anthropic phenomena occurring at the surface of the Earth [Caloz and Collet, 2001]. In the early years of remote sensing, such a crucial task has ever been carried out via photo-interpretation: aerial photographs were transformed into maps by human experts capable of recognizing forms, objects, and textures on the ground. With the advent of digital *image processing* techniques [Gonzalez and Woods, 2002], the automation of this task is now a reality. The knowledge of the analyst is complemented by the potential offered by computer programs that sequentially perform the steps needed to obtain the desired output. Expert advice is still decisive as it is required to guide the information extraction process and to assess the quality of the final product.

Land-cover classification

Techniques for tasks such as thematic classification, physical parameter retrieval via regression, data fusion, unmixing or target detection are nowadays widely investigated to derive useful georeferenced information from remotely sensed imagery [Bioucas-Dias et al., 2013]. Although many other types of spatial layers can be obtained (maps of mineral abundances, soil moisture in fields, chlorophyll content of vegetated areas, salinity of the oceans, etc.), one of the primary interests of remote sensing imagery lies in the possibility to (semi-)automatically determine an associated ground-cover for all the pixels in the scene via classification. A suitable thematic partition of the image into land-cover classes is deemed a highly valuable Earth observation product in many fields. For instance, such maps are key for urban planning, precision agriculture, forestry and land management, post-catastrophe assessment, to name a few applications.

Classification in remote sensing

Taking advantage of methods developed in the research fields of statistics, signal processing and *machine learning* (see Chapter 3), a large number of classification techniques have been successfully adapted and applied to the analysis of remote sensing data [Mather and Tso, 2003]. The starting point for this exercise is the spectral signature of the pixels recored by the sensor. Based on the vector of observations (typically DNs or reflectances), all of the procedures resort to statistical measures of distance (e. g. Euclidean or Mahalanobis distances) or similarity (e. g. Gaussian or linear kernels) to eventually determine the thematic class label of the pixels. In the following, we briefly review the main families of remote sensing image classification methods.

2.4.2 *Overview of the classification approaches by learning paradigm*

Unsupervised learning

On the one hand, *unsupervised* learning methods, also referred to as *clustering* methods, exclusively utilize the spectral signature to automatically group similar pixels into clusters. This type of techniques does not require labeled pixels: no supervision by the user is needed in the training stage. Such an approach is particularly suited when a rapid mapping is needed, for example in a change detection scenario [Bruzzone and Prieto, 2000, Bruzzone and Cossu, 2003, Volpi et al., 2012b].

Supervised learning

On the other hand, *supervised* learning methods require a set of examples with the associated class labels to be fed to the model for training. In this crucial phase, the model learns the relation between the spectral signature and the thematic class, a rule which will then be used to predict the land-cover at new locations in the prediction phase. After the early works using the parametric *Maximum Likelihood* classifier [Strahler, 1980], more recently, the remote sensing community has put the focus on powerful non-parametric machine learning methods such as neural networks [Benediktsson et al., 1990], *Support Vector Machine* (SVM) and kernel methods in general [Melgani and Bruzzone, 2004, Camps-Valls and Bruzzone, 2005], random forests [Pal, 2005], Gaussian processes [Bazi and Melgani, 2010], etc.

Semisupervised learning

A hybrid category is that of *semisupervised* learning methods, strategies whereby the unlabeled data, usually available in large quantities, are leveraged to define more robustly the class boundaries of the problem at hand. One resorts to such techniques when the labeled samples are scarce and the consideration of the underlying structure of the data provides help in regularizing the solution. Examples of such methods can be found in Bruzzone et al. [2006], Camps-Valls et al. [2007], Tuia and Camps-Valls [2009].

2.4.3 *Extensions of the classic paradigms*

Active Learning

Within the supervised learning paradigm, we find the field of *Active Learning* (AL), a booming research topic in the remote sensing community in the recent years. The underlying principle consists of the incremental addition of new samples to the training set. A dedicated search strategy is devised such that the selected pixels, after manual labeling by the operator, will maximally improve the classification accuracy (compared to a random or stratified sampling of the image). Thorough reviews of the various sampling schemes proposed for the analysis of remotely sensed images can be found in Tuia et al. [2011b], Crawford et al. [2013].

Spatial-spectral classification

Parallel to the above developments, the inclusion of spatial information in the thematic classification process has proven highly valuable [Wang et al., 1983]. Integrating a description of the arrangement of pixel values in the geographical space allows to produce smoother maps, with a higher spatial coherence, especially when considering VHR images [Tuia et al., 2009a]. These approaches are based on segmentation [Huang and Zhang, 2008], Markov random fields [Jhung and Swain, 1996, Moser et al., 2013], texture extracted through the gray level co-occurrence matrix [Haralick et al., 1973], mathematical morphology [Pesaresi and Benediktsson, 2001] and, more recently, morphological attribute filters [Dalla Mura et al., 2010].

2.5 WORKING WITH MULTIPLE IMAGES

2.5.1 *Model portability*

Temporal and spatial large-scale problems

Large-scale Earth observation problems are generally tackled by the practitioner by making use of multiple remotely sensed images. On the one hand, the mentioned scale can be temporal, with the analysis of time series of images of the same region to monitor the evolution of the land-cover [Jonsson and Eklundh, 2002] or with dedicated change detection studies [Coppin and Bauer, 1996]. On the other hand, the scale can be spatial, with regional or continental land-cover mapping efforts [Woodcock et al., 2001, Knorn et al., 2009]. The common requirement of all these large-scale applications is that the employed images bear similar radiometric characteristics. This means that the same land-covers/objects appear with comparable values on different images. However, the fact of being forced to resort to a large number of different acquisitions to be jointly analyzed makes this requirement hardly met in practice, which is exactly the situation this Thesis considers.

Model portability: from signature extension toward DA

In the context of thematic classification, the initial efforts by the community to answer these questions were undertaken in the area of remote sensing named *signature extension* [Olthof et al., 2005] (more details in Section 5.2.1). Such a research field, by taking advantage of the latest advances in statistics and machine learning, has recently evolved to a more mature and specific discipline now usually referred to as *Domain Adaptation* (DA) [Pan and Yang, 2010] (see Chapter 4). The problem that both disciplines aim to solve consists in modeling and extending the relation between spectral signatures and thematic classes collected over several scenes, reaching thus an adequate land-cover classification *model portability*. The particular case of *sample selection bias* considers the extension to the whole scene of models based on localized ground truth data sampled on a small portion of it. By matching the pixel signatures observed on a *source image* (or part of it) where we have labels to those of a *target image* where (usually) we do not, classifiers trained on the first can be used to accurately predict the classes of the second [Bruzzone and Marconcini, 2010].

2.5.2 *The issues*

Factors causing radiometric differences

Hereafter we present an overview of the factors limiting the portability of classifiers across multiple or vast acquisitions. In general, heavy radiometric differences usually exist between images taken over spatially separate regions at different time instants, even if the acquisition is carried out by the same Earth observation system. To a lesser extent, these effects also concern single images having a large extent over complex landscapes or with a heterogeneous nature of the land-cover. The main factors affecting the radiometry of the images are the following.

- Changes in atmospheric conditions (composition of the atmosphere, cloud cover, haze, etc.) [Schowengerdt, 2007].
- Differences in illumination (solar elevation angle depending on season and time of the acquisition) [Schowengerdt, 2007].
- Topography controlling terrain shading [Teillet et al., 1982].
- Seasonal variations affecting the phenology of vegetation [Reed et al., 1994].
- Changes in the acquisition geometry [Longbotham et al., 2012a] (studied in more detail in Chapter 8 for VHR images):
  - longer optical depth of the atmosphere at large off-nadir angles (low satellite elevation angles) leading to an increased up-scattered path radiance due to Rayleigh scattering,
  - varying angular distribution of the reflectance (small-scale BRDF effects),
  - solar observational cross-section effects responsible for changes in the reflectance of the objects with non-flat surfaces (e. g. pitched roofs, trees),
  - solar forward and backward scattering regimes (determined by satellite and sun positions) affecting the last two points.

2.5.3 *Radiometric normalization*

Radiometric needs

In the case of thematic classification involving multiple acquisitions, a preprocessing to normalize the images is needed. For instance, a blending of the set of images via mosaicking procedures [Homer et al., 1997] is desirable in order to obtain a vast surface coverage with one single large composite image having homogeneous characteristics. All the same, in change detection applications, most methodologies require comparable radiometries for the bi-temporal images to be profitably jointly analyzed. The possible solutions to make the images more radiometrically similar to each other, or to adjust them over their whole extent, can be classified into two categories: *absolute* or *relative* radiometric normalization strategies.

2.5.3.1 *Absolute normalization strategies*

Radiometric calibration

In the first category, we find traditional physically-based *radiometric calibration* approaches [Schowengerdt, 2007]. These procedures are targeted at compensating atmospheric, solar and topographic effects with the ultimate purpose of retrieving the original surface reflectance $\rho(\lambda)$ of the materials. The calibration is a three-stage sequence. The first level of calibration entails the conversion of the raw DNs to the original at-sensor radiance values. This step can be accomplished by knowing the sensor *gain* and *offset* parameters, specific to each acquisition and to each spectral channel. The second level consists of a transformation retrieving the corresponding radiance at the Earth's surface, the surface radiance L_λ.

Atmospheric compensation

Such a process is generally referred to as *atmospheric compensation*. By assuming the term L_λ^{sd} as equal to zero (no surface-reflected skylight), Eq. (2.1) of page 13 turns into

$$L_\lambda^s = L_\lambda^{su} + L_\lambda^{sp} = \tau_v(\lambda) L_\lambda + L_\lambda^{sp} \ . \tag{2.5}$$

Solving for L_λ we obtain

$$L_\lambda = \frac{L_\lambda^s - L_\lambda^{sp}}{\tau_v(\lambda)} \ . \tag{2.6}$$

In practice, if considering a unit view path transmittance $\tau_v(\lambda)$, this reduces to estimating the upwelling path radiance L_λ^{sp} induced by the atmosphere. Well-known approaches such as *dark object subtraction* [Chavez, 1988] actually utilize such a simplification to correct the signal by subtracting from every pixel the radiance measured over dark regions as deep lakes or heavily shadowed areas. The third and final step needed to derive $\rho(\lambda)$ is the solar and topographic compensation. Such a conversion transforms the previously computed surface radiance L_λ taking into account the solar path transmittance $\tau_s(\lambda)$, the exo-atmospheric solar spectral irradiance E_λ^0 and the incident angle $\theta(x, y)$. This last term is calculated based on the position of the sun and the topography, which is subject to the availability of a *Digital Elevation Model* (DEM). A sophisticated yet widely adopted algorithm allowing the calculation of the $\rho(\lambda)$ values is represented by *Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes* (FLAASH) [Cooley et al., 2002]. This procedure is essentially aimed at adjusting the distortions caused by the atmosphere. The compensation technique is implemented in a software package based on the MODTRAN program [Berk et al., 1987] and is valid only in the visible through SWIR spectrum (up to 3 μm). It is largely automatized but is highly dependent on numerous parameters of the atmosphere (water vapor content, aerosol optical thickness, etc.). Once $\rho(\lambda)$ is recovered, being this an inherent property of the materials at the surface of the Earth, the quantity grants more robust inter-image comparisons and analyses (portability of land-cover classifiers across acquisitions, change detection, etc.). In general, we draw the attention to the fact that effective compensation approaches are very demanding processes in terms of prior knowledge required, in particular for the atmospheric correction step (ancillary information and parameters required by physical models). Moreover, issues such as BRDF effects remain unaccounted for with this type of transformation. This is critical as these angular phenomena are emphasized in the acquisitions carried out by the latest VHR systems.

Comparisons

For a comparison of absolute and relative atmospheric compensation approaches applied to moderate resolution imagery we refer the reader to Song et al. [2001]. Instead, concerning VHR images, Pacifici et al. [2014] investigate the suitability of both raw DNs and surface reflectance data when undertaking change detection or multitemporal classification studies. By examining a long temporal series of acquisitions over the same scene, the authors point out the importance of resorting to physical quantities guaranteeing an increased invariance to changes in viewing geometry or in illumination and atmospheric conditions.

2.5.3.2 *Relative normalization strategies*

A statistical approach

On the other hand, relative normalization strategies are founded on statistical approaches. They are based on comparisons among images so that each acquisition is modified with respect to a reference image. One frequent requirement for these approaches is that the sensor having collected the images should remain the same, even though cross-sensor techniques have been developed, especially in change detection studies [Nielsen et al., 1998].

Pseudo-invariant features

For instance, regression analyses relying on radiometrically invariant objects (called *pseudo-invariant features*) such as man-made structures have often been employed to obtain a scene-to-scene radiometric normalization. Examples of this line of work can be found in Schott et al. [1988], Canty et al. [2004].

Classic histogram matching

Alternatively, rather simple image processing techniques such as *Histogram Matching* (HM) [Gonzalez and Woods, 2002] can be adopted. With this non-linear transform, the shape of the cumulative histogram of the image of interest is matched, band by band, to that of a reference image. The univariate *cumulative distribution function* (CDF) modification acts as follows. Considering a single band, for a given pixel value x_i belonging to the image to be adjusted possessing a CDF $F(X)$, the method looks for the

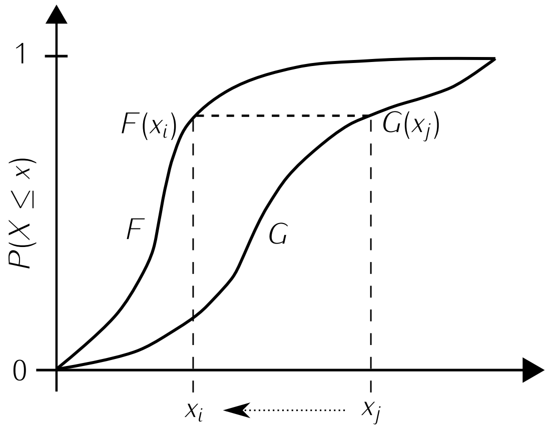


Figure 2.4: Illustration of the HM procedure.

cumulative histogram value $F(x_i)$ and finds the corresponding value in the CDF $G(X)$ of the reference image, i. e. $G(x_j) = F(x_i)$. The new pixel value x_i^*, is obtained by replacing the input pixel value x_i with x_j, thus appropriately reshaping the CDF of the image to be modified. In mathematical terms we can express this as

$$x_i^* = G^{-1}(F(x_i)) \ . \tag{2.7}$$

Figure 2.4 graphically illustrates the principle.

Extensions

The appropriate computation of the histograms to rely upon for the matching process can be challenging. They are usually computed by resorting to a discrete binning of the intensity values. Thus, the influence of the choice of number of bins, their width and placement can affect the final results. In response to these shortcomings, Gevers and Stokman [2004] propose to use a kernel density estimator. Recently, a more sophisticated HM technique has been proposed in Inamdar et al. [2008]. The authors present a multivariate extension of the univariate matching which accounts for the correlation between bands by matching the joint distributions of the images. With the same objective in mind, Zheng et al. [2012] introduce a procedure they call *Joint Histogram Matching* . A transformation of the images from the original color space to the *CIE Lab* space [McLaren, 1976] is applied before combining a univariate matching of the *lightness* dimension (L) of the two images with a matching of the joint 2-D histogram of the two *color planes* (a and b).

Part II

LEARNING FROM DATA WITH CHANGING DISTRIBUTIONS

3 MACHINE LEARNING

Outline: *In this Chapter, we present an overview of the fundamental concepts of machine learning and a description of the main techniques and approaches that will be used in this Thesis. First, the introductory Section 3.1 provides a definition and lists the main applications of the subject area. Afterwards, Section 3.2 will review the principal ways of learning from data and will introduce the statistical theory that is the foundation for all the predictive approaches adopted in this Thesis. In Section 3.3, we will focus on the Support Vector Machine classification technique and on the properties of the related kernel functions. Section 3.4 introduces the Linear Discriminant Analysis framework both for classification and for Feature Extraction. Section 3.5 addresses the latter topic in more details, namely by presenting two widely used techniques of unsupervised dimensionality reduction. In the end, the Dictionary Learning approach for the sparse representation of signals is examined in Section 3.6.*

3.1 INTRODUCTION

Definition

Machine learning can be thought of as an approach to learn from examples the dependencies existing in the data in order to perform a predictive task. Algorithms are designed such that the learning procedure takes place in a data-driven way: once the learning machine has been trained, it is used to predict the future *output* of the system at hand based on the related *input* samples [Cherkassky and Mulier, 2007]. Generally speaking, contrary to classic parametric methods developed in statistics, assumptions concerning data probability distributions are not required by machine learning procedures. Moreover, with these flexible methods, human prior knowledge can be integrated with more efficacy in the learning process, often resulting in a beneficial user-machine interaction. This rapidly growing research field can be placed at the interface between the disciplines of computer science and statistics. The terms *artificial intelligence*, *pattern recognition* and *data mining* also come into play when describing such a multifaceted science. Good foundations on the topic of machine learning and detailed explanations of the main families of techniques can be found in Bishop [2006], Cherkassky and Mulier [2007].

Applications

Machine learning has many real-world applications in diverse fields such as biology (biosequences analyses, gene expression, etc.), medicine (e. g. cancer diagnosis), chemistry (e. g. analytical chemistry), finance (e. g. stock market forecasting), web and text analysis (automatic translation, web pages categorization, hand-written character recognition, etc.). In the area of environmental sciences, the application of these developments concerns domains such as spatial interpolation (e. g. soil mapping), weather forecasting (e. g. radar-based nowcasting), natural hazards assessment (e. g. avalanches, landslides), etc. [Kanevski et al., 2008, 2009]. More specifically, data-driven methodologies naturally find a synergy within the study of geospatial data and remotely sensed images make no exception. Indeed, the breakthroughs occurred within the machine learning community have almost directly been put into practice to assist the practitioner in image analysis [Swain, 1972, Davis et al., 1978]. Especially in the last decade, the knowledge exchange and collaboration between these two scientific communities has flourished, leading to highly promising developments in the automated processing of remote sensing images [Camps-Valls, 2009, Camps-Valls et al., 2011, 2014, Tuia et al., 2014].

3.2 LEARNING FROM DATA

Two main families

Data-driven machine learning approaches can be mainly classified into two distinct categories: *supervised* and *unsupervised* learning. The techniques belonging to the former family aim at developing a model describing the input-output relationships existing in the data at hand based on the training set comprised of input sample-output label pairs. On the contrary, the latter represents an ensemble of approaches devised to extract information about the process having generated the data by solely resorting to the input samples. For the purposes of this Thesis, in the following we will consider only the supervised learning paradigm.

3.2.1 *Supervised learning*

Notation

Formally, supervised machine learning seeks relations between an input space $\mathcal{X} \in \mathbb{R}^d$ and an output space $\mathcal{Y} \in \mathbb{R}$. To this end, a *training set* $D = \{X, Y\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ composed of n labeled data samples is available to the system. Each one of these samples is described by a d-dimensional input vector $\boldsymbol{x}$ and presents a related known output y, the label. Such sample pairs are drawn from a given unknown joint probability distribution $P(X, Y)$ of variables X and labels Y. The set of input variables X will also be referred to simply as a *dataset*, whereas the above-defined D will denote more precisely a *labeled dataset*. In this Thesis, a notation using matrices will often be adopted, as many equations involving matrix calculus will appear. In such cases, a $n \times d$ *data matrix* $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ composed of the n column vectors $\boldsymbol{x}_i$ of length d belonging to dataset X will be used to represent the available training set.

The supervised approach

Starting from the input vector $\boldsymbol{x}$, the goal is find a predictive function $f(\boldsymbol{x})$ linking the input space $\mathcal{X}$ to the output space $\mathcal{Y}$ to correctly predict the associated y value. Once the appropriate model is learned, the prediction on new data takes places as follows. For each sample $\boldsymbol{x}_{\text{test}}$ belonging to an unseen *test set*, which is also following $P(X, Y)$, the machine provides a prediction $y^* = f(\boldsymbol{x}_{\text{test}})$ that can be compared to the corresponding actual label y_{test}. In Chapter 4 we will examine the issues arising in case of a different probability distributions governing the training and test data. As previously remarked, the scope of this Thesis is to develop strategies robust to this shift in probability distributions.

Regression vs. classification

Concerning the output space, the type of value of y defines the task with which we are coping. On the one hand, in *regression* problems, the output is a real value $y \in \mathbb{R}$. On the other hand, in *classification* problems, output values are discrete class labels, i. e. $y \in \mathbb{Z}$. In this case, we make the distinction between binary classification tasks usually coded with $y \in \{-1, +1\}$ and multi-class classification tasks with $y \in \mathcal{C} = \{1, 2, \ldots, c\}$, a set of c classes. A list of the main instances of supervised learning includes SVMs (see Section 3.3), neural networks, linear regression, maximum likelihood classifiers (see Section 3.4), logistic regression, decision trees and random forests, nearest neighbors, etc. [Duda et al., 2001, Cherkassky and Mulier, 2007].

3.2.2 *Statistical Learning Theory*

Framework for predictive learning

Within the field of machine learning, *Statistical Learning Theory* [Vapnik, 1998], also known as *Vapnik-Chervonenkis theory*, provides a suitable framework for predictive learning. The ultimate objective consists in the definition of appropriate models following a tradeoff between their ability to honor the available information and their complexity. In supervised learning, the function f performing the prediction can be chosen from a set of functions $\mathcal{F} = \{f(\boldsymbol{x}, \theta), \theta \in \Theta\}$, where θ represents a set of *hyper-parameters* selected from the space Θ.

Expected risk

A criterion is then required to enable us to evaluate the goodness of the choice of such a function, i. e. its similarity to the unknown target function that depicts the actual input-output dependencies. According to Vapnik's concepts, the following risk functional, called the *expected risk*, answers this need [Cherkassky and Mulier, 2007]:

$$R_{\text{exp}}(\theta) = \int L(y, f(\boldsymbol{x}, \theta))p(\boldsymbol{x}, y) \; \mathrm{d}\boldsymbol{x}\mathrm{d}y \;, \tag{3.1}$$

where the term $L(y, f(\boldsymbol{x}, \theta))$ is a task-defined loss function. The purpose of a learning algorithm is to minimize this expected average loss, keeping the risk as low as possible. Focusing on the supervised classification problem, the type of learning we will be concerned with throughout this Thesis, let us introduce the most widely employed loss function, the 0-1 loss:

$$L(y, f(\boldsymbol{x}, \theta)) = \begin{cases} 0 & \text{if } f(\boldsymbol{x}, \theta) = y \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

For this loss function, the resulting expected risk is nothing but the probability of a classification error.

Empirical risk

In practice, the *probability density function* (PDF) $p(\boldsymbol{x}, y)$ appearing in (3.1) is often unknown. The only available input-output pairs are those of the finite set of examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ the model has to rely on for training. Therefore, we approximate the theoretical risk functional with the *empirical risk* computed on the training examples as

$$R_{\text{emp}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i, \theta)) \ . \tag{3.3}$$

The minimization of this function, the *empirical risk minimization*, is then carried out to choose the best set of hyper-parameters θ. It is worth noting that as the sample size goes to infinity ($n \to \infty$), the empirical risk $R_{\text{emp}}(\theta)$ converges to the true risk $R_{\text{exp}}(\theta)$.

Structural risk

However, the fact that $R_{\text{emp}}(\theta)$ refers to the performance of the model in classifying the finite training data motivates the need for an additional term also considering the ability to extend the learned relationships to unobserved new data, the test set. The notion of *structural risk minimization* is thus introduced. Essentially, the idea is to place an upper bound for the expected risk $R_{\text{exp}}(\theta)$ of (3.1) defined as the sum of the empirical risk $R_{\text{emp}}(\theta)$ and a defined confidence interval. Mathematically, we have

$$R_{\text{exp}}(\theta) \leq R_{\text{emp}}(\theta) + \Omega(n, h) \ , \tag{3.4}$$

with the confidence interval $\Omega(n, h)$ depending on the number of training samples n and the *Vapnik-Chervonenkis* (VC) *dimension* h of the class of functions (e. g. linear, quadratic) employed [Vapnik, 1998].

Vapnik-Chervonenkis dimension

For a binary classification problem, the quantity h is the maximum number of samples for which a label-consistent partitioning of the data points can be found using the class of functions at hand, i. e. their capacity. Since more complex decision functions allow for more flexible partitions, the value h can be interpreted as a proxy for the complexity of the function. For instance, a two-dimensional training data set consisting of 3 samples can always be partitioned with a linear function, no matter the labeling of the points. Linear decision functions in $\mathbb{R}^d$ of the form $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$, where $\boldsymbol{w}$ is a d-dimensional vector of variable weights and b the associated bias, possess a VC dimension of $d + 1$. As an extreme comparison, for the class of functions $f(\boldsymbol{x}) = b \sin(\boldsymbol{w}^\top \boldsymbol{x})$ the quantity h is equal to infinity (for a sufficiently large $||\boldsymbol{w}||$), i. e. the model allows the separation of every possible configuration of training points. The expected risk is kept to a minimum when the confidence interval Ω is small. Such a situation is reached with a low h/n ratio. In fact, a complicated function possessing a large VC dimension h will perfectly fit a small number of training samples n but will still result in a large expected risk by its high complexity. This situation will likely lead to an important generalization error on new data. To summarize, the structural risk minimization principle provides a theoretical framework for achieving the optimal tradeoff between the classification accuracy on training data and the capacity of the set of functions selected.

3.2.3 *Model selection and model assessment*

Model selection

When concretely applying a supervised learning approach there are several practical considerations that need to be respected. First, the *model selection* step is crucial. A learning machine that reached a very low empirical risk (training error) by perfectly fitting noisy or non-representative training data, is said to be *overfitting* the data (in opposition to a too simple model giving rise to the situation called *underfitting*). Overfitting will result in a poor generalization ability of the system when dealing with new data.

Tuning of the hyper-parameters

Hence, after having fixed the class of functions, it is required that the selection of the optimal set of hyper-parameters θ defining the model is carried out on an independent dataset (different from the training set). Note that in this more theoretical Section 3.2, to be consistent with the terminology defined in Statistical Learning Theory, the term hyper-parameters is employed to distinguish these global tuning parameters (e. g. C or σ for the SVM, see next Section 3.3) from the actual parameters of the learning machine that are obtained by the algorithm after internally solving an optimization problem (e. g. α coefficients of the SVM). Nonetheless, in the rest of the manuscript we will refer to hyper-parameters simply as the *parameters* of the model.

Cross-validation

Starting with the training set alone, a common and easy solution to simulate the availability of a separate set of samples consists in tuning the hyper-parameters via cross-validation procedures (K-fold or leave-one-out). In classification for instance, predictions of class membership are performed on a held-out subset of the training data, the validation set, by using the rest of the set of samples to train the model. We purposely ignore the known class labels in the held-out set so that the agreement between the true and predicted class assignments can be checked. This procedure is repeated by partitioning the training set as many times as needed to test all the training-validation combinations. A grid-search over the space spanned by Θ allows then the user to determine the best hyper-parameters for the classification task. Such a cross-validation process, nonetheless, remains strongly dependent on the examples provided to the learning machine for training (see sample selection bias issues addressed in Chapter 4).

Model assessment

Finally, the *model assessment* step is needed to assess the generalization error of the selected model. To this end, an independent test set should be used, when at all possible, to assess the true performance of the model. Indeed, it is not fair to report the best performances observed during the previously executed cross-validation as a measure of success because the learning machine is biased favorably to this data (hyper-parameters perfectly tuned for this set) [Kanevski et al., 2008].

Metrics

In Appendix A, the reader will find a description of the most widely used metrics to assess the quality of thematic maps produced by supervised classification of remote sensing images. These measures can be used to evaluate the performances during both the model selection and model assessment phases.

3.3 SUPPORT VECTOR MACHINES AND KERNELS

In this Section we will present one of the main supervised learning systems used in this Thesis: the SVM classifier. The technique is a *large margin classifier* belonging to the family of *kernel methods* [Shawe-Taylor and Cristianini, 2004] and rigorously adheres to the guidelines provided by Statistical Learning Theory discussed in Section 3.2.2.

3.3.1 *Large margin linear classifier*

Optimal separating hyperplanes

We will examine here the reasons why a linear decision function can optimally be used as a foundation for the classification task. In a d-dimensional space, a set of training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ belonging to two categories $y_i = +1$ or $y_i = -1$ can be effectively partitioned by placing a hyperplane $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$. The input vector $\boldsymbol{x} \in \mathbb{R}^d$ describing each sample is multiplied by a weighting vector $\boldsymbol{w}$ which needs to be retrieved along with the offsetting scalar b. The new data points are labeled following the sign of the function $f(\boldsymbol{x})$: they are classified either in the positive class ($y_i^* = +1$) if $f(\boldsymbol{x}) > 0$ or, otherwise, in the negative class ($y_i^* = -1$) if $f(\boldsymbol{x}) < 0$. On the training dataset, the decision function $f(\boldsymbol{x})$ should respect

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall\, i\ . \tag{3.5}$$

The *slack variables* ξ_i allow noisy training samples to lie inside the region between $f(\boldsymbol{x}) = +1$ and $f(\boldsymbol{x}) = -1$ referred to as the *margin*. In order to keep low the empirical error of (3.3) one should, of course, force the algorithm to assign non-zero ξ_i values to as few as possible of the training samples (see Eq. (3.6)). This formulation is referred to as *soft margin* SVM and provides more flexibility with respect the a *hard margin* principle where all the training data points are forced to be outside the margin, that is when $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1\ \forall\, i$ holds.

Support vectors

Besides the few training samples lying in the margin, most of the training points should bear a decision function $f(\boldsymbol{x}_i) > +1$ if $y_i = +1$ and $f(\boldsymbol{x}_i) < -1$ if $y_i = -1$. Meanwhile, the points in correspondence of whom $f(\boldsymbol{x})$ takes the exact values $+1$ or -1 are called *Support Vectors* (SVs).

Large margin

The ultimate goal of a supervised classifier is to suitably generalize the rules learned from the training data to any new set of instances that has to be classified (the test set). The situation in which most of the new data points will likely be correctly labeled is reached by setting the largest possible margin. Since the margin has a width of $\varrho = 2/\,||\boldsymbol{w}||$, the search for this optimal separating hyperplane can be guided by the minimization of $||\boldsymbol{w}||$. Moreover, such a minimization problem is theoretically justified by the principles of the Statistical Learning Theory [Vapnik, 1998]. Figure 3.1 pictures the main elements defining the soft margin SVM.

Primal formulation

The algorithm behind SVMs provides an efficient solution to maximize ϱ while respecting the constraints in (3.5). These two objectives can be combined in the following minimization task, i. e. the *primal formulation* of

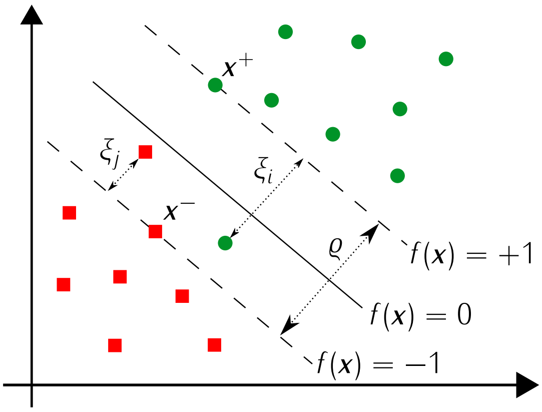


Figure 3.1: Illustration of the soft margin SVM principle. Samples $\boldsymbol{x}$ with class labels $y = +1$ appear with **green circles** whereas samples with class labels $y = -1$ appear with **red squares**. Data points denoted with $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ constitute the SVs of the positive and negative classes defining the hyperplane. In this example, slack variables ξ_i and ξ_j are assigned to positive and negative noisy samples lying beyond the margin boundary of their class.

the SVM problem, which privileges simple functions with large margins (left term) and tries to commit as little errors on the training set as possible (right term):

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right\} \tag{3.6}$$

$$\text{s.t.} \quad \xi_i \geq 0 \,, \tag{3.7}$$

$$y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i \quad \forall i \,. \tag{3.8}$$

The tradeoff constant C (the *cost* or *penalty parameter*) allows the user to control the number of errors allowed during the training phase. A value of C that is too large implies almost no training errors, forcing a highly complex model eventually incurring in the risk of overfitting. Conversely, a too small C permits many misclassified training samples, leading to an over-simplistic model. The above formulation involving the constraint (3.8) is associated with a *hinge loss* $L(y, f(\boldsymbol{x}, \theta)) = \max(0, 1 - yf(\boldsymbol{x}, \theta))$ which differs from the classic binary 0-1 loss of (3.2). Unlike the latter that merely looks for misclassifications, the hinge loss depends on how far the samples are from the hyperplane. Being thus a continuous function, it ensures an optimal and tractable solution for the SVM problem.

Dual formulation

To this end, after introducing *Lagrange multipliers* $\alpha_i \geq 0$ (dual sample weights) associated with each training sample $\boldsymbol{x}_i$, the *dual formulation* is derived as

$$\max_{\alpha} \quad \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j \right\} \tag{3.9}$$

$$\text{s. t.} \quad \sum_{i=1}^{n} y_i \alpha_i = 0 \ , \tag{3.10}$$

$$0 \leq \alpha_i \leq C \quad \forall i \ . \tag{3.11}$$

A comprehensive description of these steps can be found in Schölkopf and Smola [2002], Cristianini and Shawe-Taylor [2000], Hastie et al. [2009].

Linear SVM decision function

After having solved this convex *quadratic programming* problem yielding a unique solution, the final SVM decision function for a generic unseen vector $\boldsymbol{x}$ can be formulated as

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \alpha_i \boldsymbol{x}_i^\top \boldsymbol{x} + b \ . \tag{3.12}$$

When facing a binary classification task, the predicted class label y_i^* ($+1$ or -1) is simply assigned following the sign of (3.12). In a multi-class case ($c > 2$), the solution consists in combining several binary classifiers with either a *one-vs-all* approach (c binary SVMs separating a given class from all the rest) or a *one-vs-one* approach ($c(c-1)/2$ binary SVMs coping with two classes at a time) [Schölkopf and Smola, 2002]. When adopting the first strategy, one assigns the sample to the class which has the largest $f(\boldsymbol{x})$ value, whereas with the second strategy a majority vote is used to select the winning class.

Relevance of the SVs

The main output of the SVM training procedure are the dual coefficients α_i controlling the definition of the decision function. From Eq. (3.12) we realize that these coefficients are nothing but weights given to each training sample $\boldsymbol{x}_i$. Only a small fraction of them receives a non-zero α_i, implying that solely an exclusive subset of the initial training set is actually contributing in the evaluation of the decision function for a given new point $\boldsymbol{x}$. These highly informative samples are the same SVs already mentioned above for which $y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$ holds. Note that the ratio of SVs to the total number of training points carries an important meaning: the higher the ratio, the more the model is fitted to the training data. In fact, in such a situation many SVs contribute to the final SVM solution, leading to a complex prediction model. Furthermore, let us recall that the upper bound for the α_i is set by the penalty parameter C, so that $0 \leq \alpha_i \leq C$, $\forall i$. Such a property will be of interest in Chapter 6.

3.3.2 *Non-linear extension: the kernel trick*

The principle

Hereafter, we will build on the linear SVM presented above by addressing the developments enabling non-linear decision functions. Indeed, when dealing with challenging datasets, the input-output relationships are seldom linear. In this situation, the two classes of interest can only be suitably discriminated by a non-linear boundary. Rather than applying complex decision functions directly on the initial data set, the intuition (Cover's theorem) consists in mapping the dataset into a space of higher dimension and then only there, on the transformed data, perform the well-known linear separation [Cover, 1965].

Kernel trick

This is possible since in Eq. (3.12), the calculation of $f(\boldsymbol{x})$ involves a dot product between the input vector $\boldsymbol{x}$ whose prediction is being computed and all the training samples $\boldsymbol{x}_i$. Therefore, by means of the so-called *kernel trick*, the idea is to substitute the dot product with a *kernel function* $K(\cdot,\cdot)$ involving the same two vectors, so that the final SVM decision function changes to

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} y_i \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{3.13}$$

Mapping via kernel functions

The function $K(\cdot,\cdot)$ implicitly carries out an implicit mapping φ to a higher-dimensional space, referred to as *reproducing kernel Hilbert space* (RKHS). As a matter of fact, it does not directly generate new vectors for the two samples in the mapped space. Instead, it concentrates on the result of the dot product involving the mapped vectors $\varphi(\boldsymbol{x}_i)$ and $\varphi(\boldsymbol{x})$, which should be equal to the output of the kernel computed with the low-dimensional vectors as inputs:

$$\boldsymbol{x}_i^\top \boldsymbol{x} \mapsto \varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}) = K(\boldsymbol{x}_i, \boldsymbol{x}) \ . \tag{3.14}$$

In machine learning, we refer to the original space as the *input space*, whereas we name the kernel-induced one the *feature space*.

3.3.3 *Kernel functions*

Valid kernel functions

A wide range of different kernel functions that can be applied, especially as the rapid developments in the field of kernel methods are continuously bringing up new variants adapted to specific problems. However, note that not every function taking two vectors as input constitutes a kernel. In fact, valid kernels have to fulfill the *Mercer's conditions* [Vapnik, 1998, Cristianini and Shawe-Taylor, 2000]. These constraints must be met for a selected function $K(\cdot,\cdot)$ to act as a kernel associated with the desired feature space. Strictly speaking, this means that the $n \times n$ *kernel matrix* $\boldsymbol{K} = \left(K_{i,j}\right)_{i,j=1}^{n} = \left(K(\boldsymbol{x}_i, \boldsymbol{x}_j)\right)_{i,j=1}^{n}$ also known as *Gram matrix*, has to be symmetric and positive semidefinite (possess non-negative eigenvalues).

Examples of kernel functions

Hereafter, we list some of the most widely used kernel functions:

- Linear kernel:
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_j \tag{3.15}$$
- Gaussian RBF kernel:
$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) \ , \quad \sigma \in \mathbb{R}^+ \tag{3.16}$$

- Laplace kernel:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right) , \quad \sigma \in \mathbb{R}^+ . \tag{3.17}$$

The first item, the *Linear kernel*, corresponds to the situation where the kernel trick has not been applied and computes a similarity based on the dot product only, i. e. the cosine between the two vectors. The second and third kernels listed, the *Gaussian Radial Basis Function* (RBF) *kernel* and the *Laplace kernel*, both consists of an exponential function with an argument involving a dissimilarity measure between vector $\boldsymbol{x}_i$ and vector $\boldsymbol{x}_j$ rescaled by a kernel width parameter σ. In fact, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the Euclidean distance between the examples computed in the input space. Such kernels offer an intuitive geometrical interpretation: they reflect the similarity between the samples.

Properties of kernels

Additionally, thanks to the properties of these functions, user-defined kernels can be created by multiplying or adding valid kernels since the resulting functions also respect Mercer's conditions. If $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ are valid kernels,

$$aK_1(\cdot,\cdot) + bK_2(\cdot,\cdot) \qquad \text{for} \quad a, b > 0 , \tag{3.18}$$

$$K_1(\cdot,\cdot)K_2(\cdot,\cdot) , \tag{3.19}$$

are valid kernels as well [Genton, 2002, Shawe-Taylor and Cristianini, 2004]. These properties permit the construction of *composite kernels* that may improve the classification performance of the SVM [Camps-Valls and Bruzzone, 2005].

3.4 LINEAR DISCRIMINANT ANALYSIS

3.4.1 *Maximum Likelihood classifier*

Bayesian decision theory

Another supervised classifier we will often resort to in this dissertation for its simplicity (no hyper-parameters to tune) and ease of application (rapidly computed) is the *Linear Discriminant Analysis* (LDA) [Fukunaga, 1990]. Although with a strict categorization this approach would not be considered as belonging to machine learning, it is a predictive learning tool as well. Such a model, also known as the *Maximum Likelihood classifier*, is a parametric classifier (hypothesis of data normality) based on Bayesian decision theory. In this context, it is suggested that with the knowledge of class posterior probabilities for the samples at hand, an optimal classification can be obtained. Considering Bayes' theorem, a sample $\boldsymbol{x}$ should be assigned to the class cl with the largest posterior probability

$$P(cl|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|cl)P(cl)}{p(\boldsymbol{x})} , \tag{3.20}$$

where $P(cl)$ is the prior probability of each class $cl = 1, \dots, c$, whereas $p(\boldsymbol{x}|cl)$ and $p(\boldsymbol{x})$ are the class-conditional and marginal PDFs of sample $\boldsymbol{x}$, respectively.

Discriminant function

If we suppose that the samples of each class cl are drawn from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{cl}$ and covariance matrix Σ_{cl} we have

$$p(\boldsymbol{x}|cl) = \frac{1}{(2\pi)^{d/2}|\Sigma_{cl}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{cl})^\top \Sigma_{cl}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{cl})\right) . \quad (3.21)$$

Substituting (3.21) in (3.20), taking its natural logarithm and then dropping the terms that are independent of cl, we obtain the following discriminant function

$$f_{cl}(\boldsymbol{x}) = \ln(P(cl)) - \frac{1}{2}\ln|\Sigma_{cl}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{cl})^\top \Sigma_{cl}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{cl}) . \quad (3.22)$$

Finally, the predicted class label y^* for sample $\boldsymbol{x}$ is attributed according to

$$y^* = \arg\max_{cl} f_{cl}(\boldsymbol{x}) . \quad (3.23)$$

Maximum likelihood parameter estimate

Mean vectors $\boldsymbol{\mu}_{cl}$ and covariance matrices Σ_{cl} for each class can be derived from the training data using classic maximum likelihood estimates. The class mean vector is estimated as $\hat{\boldsymbol{\mu}}_{cl} = \bar{\boldsymbol{x}}_{cl}$, corresponding thus to the class sample mean vector. The $d \times d$ covariance matrix is estimated with $\hat{\Sigma}_{cl} = \frac{1}{n_{cl}-1}\boldsymbol{S}_{cl}$, where n_{cl} is the number of training samples $\boldsymbol{x}_{cl_i}$ belonging to class cl and

$$\boldsymbol{S}_{cl} = \sum_{i=1}^{n_{cl}} (\boldsymbol{x}_{cl_i} - \bar{\boldsymbol{x}}_{cl})(\boldsymbol{x}_{cl_i} - \bar{\boldsymbol{x}}_{cl})^\top \quad (3.24)$$

is the *scatter matrix* of class cl. Prior probabilities are obtained as $P(cl) = n_{cl}/n$.

Linear vs. Quadratic vs. Naive Bayes

In the special case where a common *pooled within-class covariance* matrix Σ_W is assumed for all the classes, i. e. $\hat{\Sigma}_{cl} = \Sigma_W = \frac{1}{n-c}\sum_{cl=1}^{c} \boldsymbol{S}_{cl}\ \forall cl$, we refer to this method as LDA. Conversely, the version with separate class-specific $\hat{\Sigma}_{cl}$ gives rise to *Quadratic Discriminant Analysis*. As the name suggests, the former yields linear class boundaries, while the latter provides non-linear, quadratic boundaries. Furthermore, a special case of LDA is represented by the *Naive Bayes classifier*, which arises if we assume that all the variables are independent, that is the class-conditional densities are computed based on a pooled covariance matrix that is diagonal.

3.4.2 *Linear Discriminant Analysis for Feature Extraction*

Fisher Discriminant Analysis

The method presented above for a classification problem can also be considered from the perspective of *Feature Extraction* (FE) (see next Section 3.5). Indeed, this corresponding formulation originates from the *Fisher's linear discriminant* [Fisher, 1936], a technique developed for two-class problems and without making the assumption of normally distributed classes. The underlying principle consists in seeking the directions in our initial input space that ensure the best discrimination of the classes. More precisely, in a multi-class case, the ideal purpose is to obtain a new subspace in which the centroids (sample means) of the classes are the most spread, while the variance of the data points within the classes is the smallest.

Optimization problem

Mathematically, these directions are given by the vectors $\boldsymbol{u}$ maximizing the ratio of the between-class scatter to the within-class scatter in the projected space, i. e. maximizing the following *Rayleigh quotient*

$$\arg\max_{\boldsymbol{u}} \frac{\boldsymbol{u}^\top \boldsymbol{S}_B \boldsymbol{u}}{\boldsymbol{u}^\top \boldsymbol{S}_W \boldsymbol{u}} \ . \tag{3.25}$$

Matrices $\boldsymbol{S}_B$ and $\boldsymbol{S}_W$ are the *between-class scatter matrix* and the *within-class scatter matrix* in the input space and are computed as

$$\boldsymbol{S}_B = \sum_{cl=1}^{c} n_{cl}(\bar{\boldsymbol{x}}_{cl} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_{cl} - \bar{\boldsymbol{x}})^\top \tag{3.26}$$

$$\boldsymbol{S}_W = \sum_{cl=1}^{c} \boldsymbol{S}_{cl} \ , \tag{3.27}$$

with $\bar{\boldsymbol{x}}$ representing the total sample mean vector. We find projection vectors $\boldsymbol{u}$ by solving the generalized eigenvalue problem

$$\boldsymbol{S}_B \boldsymbol{u} = \rho \boldsymbol{S}_W \boldsymbol{u} \ , \tag{3.28}$$

where $\{\boldsymbol{u}_i\}_{i=1}^d$ are the *eigenvectors* defining the projection and $\{\rho_i\}_{i=1}^d$ are the associated *eigenvalues*. Let us remark that the number of non-zero eigenvalues is actually $c-1$ at most and this bounds the maximum number of extracted features representing the directions of greatest class separability.

Projection of new samples

Eventually, the projection of some test data points arranged in a $n_{\text{test}} \times d$ data matrix $\boldsymbol{X}_{\text{test}}$ (e. g. all the pixels of a remote sensing image) onto these newly extracted *discriminant components* is carried out through

$$\boldsymbol{X}^*_{\text{test}} = \boldsymbol{X}_{\text{test}} \boldsymbol{U} \ , \tag{3.29}$$

where $\boldsymbol{U}$ is the $d \times (c-1)$ matrix constituted by the $c-1$ eigenvectors $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{c-1}]$ of (3.28).

For more complete developments and for an in-depth discussion of the connections between the classification and FE frameworks of LDA we refer the reader to Duda et al. [2001] and to Hastie et al. [2009]. A kernel-based extension of this discriminant dimensionality reduction method called *Generalized Discriminant Analysis* (GDA) has been proposed by Baudat and Anouar [2000]. Such an implementation allows a non-linear supervised FE that can cope with multi-class problems.

3.5 FEATURE EXTRACTION

Objectives

Many datasets encountered in machine learning and, more and more often also in remote sensing, present a very high number of variables (hundreds to thousands). When relying on a limited amount of training samples, the predictive power of a learning machine is hindered by a too large number of dimensions describing the data. Such a negative effect is called *Hughes phenomenon* or *curse of dimensionality* [Hughes, 1968]. To cope with this issue, as well as with the intrinsic non-linearities in the data, an option is that of dimensionality reduction via FE. The purpose is to map the original data into a space of (much) lower dimensionality while preserving their main characteristics [Arenas-García and Petersen, 2009]. In this Section we describe two key FE techniques: the linear *Principal Component Analysis* (PCA) and its non-linear kernel-based extension, the *Kernel Principal Component Analysis* (KPCA). They are considered unsupervised FE methods because they do not make use of the label information in the definition of the mapping. In contrast, the LDA technique presented in the preceding Section is a supervised approach to dimensionality reduction. We recall that in Chapter 7 we will consider FE methodologies to tackle adaptation problems.

3.5.1 *Principal Component Analysis*

The basics

The multivariate techniques known as PCA [Hotelling, 1933] allows to convert an initial set of correlated variables into a new set of linearly uncorrelated variables referred to as the *principal components*. The procedure is based on a multidimensional rotation relying on an eigenvalue decomposition of the covariance matrix of the initial data. The newly extracted principal components, besides being orthogonal to each other, also aim at keeping a maximum of the original data variance.

Primal PCA formulation

Let us consider the $n \times d$ data matrix $\boldsymbol{X}$ (with columns centered to a zero mean). The objective of classical PCA is to find the directions of maximal variance by diagonalizing the $d \times d$ covariance matrix $\Sigma = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X}$. In the primal formulation (*R-mode* analysis), this is carried out by solving the following eigenproblem:

$$\frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{u} = \rho\boldsymbol{u} \;, \tag{3.30}$$

where $\{\boldsymbol{u}_i\}_{i=1}^d$ and $\{\rho_i\}_{i=1}^d$ are the eigenvectors and the respective eigenvalues. The largest eigenvalue is associated with the eigenvector specifying the direction of greatest variability in the initial data.

Projection of new samples

The projection of some test samples $\boldsymbol{X}_{\text{test}}$ is usually done on the first m principal components ($m \ll d$). Such a mapping is obtained as

$$\boldsymbol{X}^*_{\text{test}} = \boldsymbol{X}_{\text{test}}\boldsymbol{U} \;, \tag{3.31}$$

where $\boldsymbol{U}$ in this case is a $d \times m$ matrix constituted by the first m eigenvectors $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]$ (ordered by decreasing eigenvalue).

3.5.2 *Kernel Principal Component Analysis*

The basics

Introduced by Schölkopf et al. [1998], KPCA is the non-linear extension of standard PCA. As its linear counterpart, KPCA aims at extracting a set of features enhancing the data representation in a subspace of reduced dimensionality. The extracted components are still orthogonal to each other, but, contrary to PCA, these are no more simple linear combinations of the input variables.

Dual PCA formulation

In the corresponding dual formulation (*Q-mode* analysis) of PCA eventually leading to KPCA, instead of the covariance matrix, we analyze the $n \times n$ Gram matrix $\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\top$, i. e. the kernel matrix obtained with a linear kernel. With this formulation, the PCA eigenproblem becomes

$$\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{v} = \rho \boldsymbol{v} \ , \tag{3.32}$$

and yields dual eigenvectors $\{\boldsymbol{v}\}_{i=1}^{n}$ and eigenvalues $\{\rho\}_{i=1}^{n}$. It is possible to show that the eigendecomposition yields the same non-zero eigenvalues ρ_i whose eigenvectors $\boldsymbol{v}_i$ are related to their primal counterparts $\boldsymbol{u}_i$ by $\boldsymbol{v}_i = \boldsymbol{X}\boldsymbol{u}_i/\sqrt{(n-1)\rho_i}$ [Nielsen and Canty, 2008].

KPCA formulation

Since in this representation a dot product between samples $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ ($\boldsymbol{X}\boldsymbol{X}^\top$ in matrix notation) comes into play, we take advantage of the previously introduced kernel trick (see Section 3.3.2) to implicitly simulate a mapping φ of the samples into a higher-dimensional RKHS. Consequently, Eq. (3.32) becomes

$$\begin{aligned} \tfrac{1}{n-1}\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})^\top \boldsymbol{v} &= \rho \boldsymbol{v} \quad \Leftrightarrow \\ \tfrac{1}{n-1}\boldsymbol{K}\boldsymbol{v} &= \rho \boldsymbol{v} \ , \end{aligned} \tag{3.33}$$

where $\boldsymbol{K}$ is the kernel matrix with elements $K_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^\top \varphi(\boldsymbol{x}_j)$. Dropping the $1/(n-1)$ factor and by using the centered kernel matrix $\tilde{\boldsymbol{K}} = \boldsymbol{H}\boldsymbol{K}\boldsymbol{H}$, with a $n \times n$ centering matrix $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^\top/n$, the final KPCA eigenvalue problem is set up as

$$\tilde{\boldsymbol{K}}\boldsymbol{v} = \rho \boldsymbol{v} \ . \tag{3.34}$$

Projection of new samples

As seen in (3.32), opposed to the primal PCA, the number of features the user is allowed to extract is bounded by the number of training samples n and not by the number of initial variables d. This is due to the fact that the eigendecomposition involves a $n \times n$ matrix (the kernel matrix). The projection of new samples $\boldsymbol{X}_{\text{test}}$ on the first m *kernel principal components* ($m \ll n$) is computed as

$$\boldsymbol{X}^*_{\text{test}} = \tilde{\boldsymbol{K}}_{\text{test}} \boldsymbol{V} \ , \tag{3.35}$$

where $\tilde{\boldsymbol{K}}_{\text{test}}$ is the $n_{\text{test}} \times n$ centered test kernel matrix between the n_{test} test samples and the n training samples and $\boldsymbol{V}$ is constituted by the first m eigenvectors $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$.

3.6 DICTIONARY LEARNING

Usefulness of sparse representations

This Section introduces the basic concepts necessary for the developments proposed in Chapter 9. We will explore a family of methods belonging to the *sparse representation* [Wright et al., 2010] framework based on *Dictionary Learning* (DL) [Aharon et al., 2005, Tosic and Frossard, 2011]. This area of research is more related to classical signal processing and its sub-field of *compressive sensing* but, as we will see in Section 3.6.2, it can provide powerful predictive tools. In general terms, the sparse representation of signals has proven to be an extremely useful tool to represent and compress high-dimensional real-world signals. Data such as audio recordings or images, for instance, present a sparse nature which immediately lends itself to compression. Sparse representations take advantage of this very fact to express a sample as a combination of a few other samples, the *atoms*, constituting a reference set called the *dictionary*. In particular, sparse representations have been successfully used in a variety of image processing tasks including image denoising, inpainting and classification [Wright et al., 2010].

3.6.1 *Learning the dictionary*

The optimization problem

Given a training data matrix $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ of n column vectors $\boldsymbol{x}$ of dimension d (the signals)[1], DL basically consists in finding a dictionary $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_i, \ldots, \boldsymbol{d}_K] \in \mathbb{R}^{d \times K}$ composed of K atoms $\boldsymbol{d}_i$ and a matrix $\boldsymbol{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_i, \ldots, \boldsymbol{c}_n] \in \mathbb{R}^{K \times n}$ composed of n vectors of *sparse codes* $\boldsymbol{c}_i$ such that $\boldsymbol{X} \approx \boldsymbol{DC}$. The search for the dictionary $\boldsymbol{D}$ and associated matrix $\boldsymbol{C}$ allowing to recover the original data $\boldsymbol{X}$ can be expressed with the following optimization problem:

$$\{\boldsymbol{D}, \boldsymbol{C}\} = \underset{\boldsymbol{D},\boldsymbol{C}}{\arg\min} \left\| \boldsymbol{X} - \boldsymbol{DC} \right\|_F^2 \quad \text{s.t.} \quad \left\| \boldsymbol{c}_i \right\|_0 \leq s \quad \forall\, i\,. \tag{3.36}$$

The operator $\|\cdot\|_0$ denotes the ℓ_0 norm counting the number of nonzero entries in a vector and s is the sparsity level, i. e. the desired number of nonzero coefficients used in the retrieval of each signal.

K-SVD

The problem in (3.36) is NP-hard and therefore only an approximate solution can be found. To this end, an approach named K-SVD [Aharon et al., 2005] can be adopted. It consists of an alternate minimization using the greedy *Orthogonal Matching Pursuit* (OMP) [Tropp and Gilbert, 2007] to find a sparse $\boldsymbol{C}$ for a given $\boldsymbol{D}$, followed by a minimization over $\boldsymbol{D}$ where each column of it, the atoms, is modified to better represent the signals in $\boldsymbol{X}$. After an appropriate number of iterations, $\boldsymbol{D}$ will be particularly suitable for the sparse representation of signals similar to those forming the training set $\boldsymbol{X}$.

[1] Please remark that the notation used in the field of sparse coding/DL differs from that normally used in machine learning where the data matrix $\boldsymbol{X}$ is transposed, i. e. of size $n \times d$. Moreover, in this Section 3.6 as well as in Chapter 9, the term "signal" is interchangeably used with the term "sample" to refer to the data vector $\boldsymbol{x}$.

3.6.2 *Dictionary-based classification*

Good classification performances have been obtained by applying classification routines based on dictionaries to a variety of problems [Kong and Wang, 2012]. In the field of remote sensing, the applications have mainly focused on hyperspectral image classification [Chen et al., 2011b, Wang et al., 2014, Li et al., 2014]. Basically, when facing a supervised classification task, two main approaches exist to tackle the problem using dictionaries.

Discriminative coefficients

On the one hand, a first class of strategies aims at guiding the learning process to make the coefficients $\boldsymbol{C}$ more discriminative. In Mairal et al. [2008], an approach adding to (3.36) the logistic regression loss function is presented, whereas in Jiang et al. [2011] the authors introduce a label consistent K-SVD, a method adding to the minimization problem a term forcing the signals belonging to the same class to have similar sparse representations. These approaches either learn the parameters of a classifier (e. g. weights of a linear predictor) in parallel during the optimization process or, once the final coefficients $\boldsymbol{C}$ have been determined, they subsequently feed these now discriminant coefficients to an external classifier.

Discriminative dictionaries

On the other hand, attention could be paid to directly make the dictionary $\boldsymbol{D}$ more discriminative [Yang et al., 2010]. To predict the class label of the test samples, such an approach effectively exploits the mentioned specialization of the dictionary to represent the training set. Considering a c-class classification task, the problem is set up as follows. By means of Eq. (3.36), one starts by learning c class-specific dictionaries $\{\boldsymbol{D}_{cl}\}_{cl=1}^{c}$, each time by leveraging exclusively the training samples of that very same class to form $\boldsymbol{X}$.

Global vs. specific coding

Once the global dictionary $\boldsymbol{D} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_c]$ made of the class-specific $\boldsymbol{D}_{cl}$ has been learned, there are two possible approaches to compute (via OMP for instance) the sparse representation, i. e. the sparse coding, of a new test sample $\boldsymbol{x}_{\text{test}}$ in terms of $\boldsymbol{D}$:

- *Global coding*: the whole $\boldsymbol{D}$ is used at once to sparsely represent $\boldsymbol{x}_{\text{test}}$ with the coefficients $\boldsymbol{c}_{\text{test}}$. A single coding process takes place and the s nonzero coefficients are shared by all the sub-dictionaries. Class-specific coefficients $\boldsymbol{c}_{\text{test},cl}$ can be retrieved as the respective portions of the global vector $\boldsymbol{c}_{\text{test}}$.

- *Specific coding*: each $\boldsymbol{D}_{cl}$ is separately used to sparsely represent $\boldsymbol{x}_{\text{test}}$ by means of $\boldsymbol{c}_{\text{test},cl}$, a class-specific vector of coefficients. These dedicated coefficients can then be concatenated to form the global $\boldsymbol{c}_{\text{test}} = [\boldsymbol{c}_{\text{test},1}^{\top}, \ldots, \boldsymbol{c}_{\text{test},c}^{\top}]^{\top}$. There are as many sparse coding processes as classes, each one using only the atoms of the sub-dictionary $\boldsymbol{D}_{cl}$ of the corresponding class.

Reconstruction error to assign the class label

In both cases, the coefficients $\boldsymbol{c}_{\text{test},cl}$ and the associated dictionaries $\boldsymbol{D}_{cl}$ of each class will be in turn used to represent a new test sample $\boldsymbol{x}_{\text{test}}$. The goodness of the representation is usually measured using the ℓ_2 norm of the reconstruction error

$$r_{cl} = \left\| \boldsymbol{x}_{\text{test}} - \boldsymbol{D}_{cl} \boldsymbol{c}_{\text{test},cl} \right\|_2 \ , \tag{3.37}$$

where a small error r_{cl} indicates a high affinity of $\boldsymbol{x}_{\text{test}}$ with class cl. The signal will be assigned to the class whose $\boldsymbol{D}_{cl}$ yields the best reconstruction, meaning that the predicted label for $\boldsymbol{x}_{\text{test}}$ is determined as

$$y^* = \arg\min_{cl} r_{cl} \ . \tag{3.38}$$

In the global coding case, it can occur that all the coefficients of a sub-dictionary receive a zero value. As a consequence, these sub-dictionaries will be unable to reconstruct the pixel at all. Moreover, the coefficient with the largest magnitude is generally dominant, as the remaining coefficients only receive a negligible weight. Thus, the test signal is almost always directly assigned to the class receiving the largest coefficient in $\boldsymbol{c}_{\text{test}}$. In fact, the sub-dictionary $\boldsymbol{D}_{cl}$ that possesses the most similar atom to $\boldsymbol{x}_{\text{test}}$ will decide the class attribution. On the contrary, the specific coding returns $s \cdot c$ active coefficients, allowing each class to reconstruct the signal to some extent. The residuals associated with the classes are therefore relatively small and comparable with each other. An analysis of these two strategies and an approach exploiting their complementarity in the context of hyperspectral classification is provided in [Marcos Gonzalez et al., 2014].

4 DOMAIN ADAPTATION

Outline: *In this fourth Chapter, we tackle the very problem that this Thesis attempts to solve: find a proper adaptation strategy to continue learning from data also when the underlying distributions change from one dataset to another. Section 4.1 outlines what are the issues arising from shifting distributions and lists the sub-fields of machine learning dealing with such problems. In Section 4.2, we define the specific notation and formalize the basic Domain Adaptation concepts used throughout the rest of the manuscript. Section 4.3 presents some measures that can be used to assess the degree of shift: both classic and recently proposed metrics of distance between probability distributions are introduced. Ultimately, Section 4.4 reviews the existing approaches to adaptation while proposing their classification into three distinct families.*

4.1 INTRODUCTION

Differences between training and test sets

Within the field of machine learning, the large majority of predictive approaches proposed up to these days relies on a widespread key assumption: the training and test datasets are drawn from the same probability distribution [Pan and Yang, 2010]. Another additional and more easily met condition required by most of the methods is that the variables describing the samples need to be the same or at least they need to be measuring the same phenomena. It is well-known that when the input space or the distributions governing the data change, classic statistical models fail at suitably generalizing the learned properties over multiple datasets. As a matter of fact, for each new dataset, the user needs to collect every time a series of training samples and build a new model from scratch. Depending on the application, such a process might be expensive or even impossible to be completed.

Solutions

A more attractive option consists in developing effective strategies to ease the *knowledge transfer* between datasets. Indeed, the ability to re-utilize pieces of information collected on a different but related set of data in further applications is highly desirable. Over the last decade, the study of such methods has become more and more popular within the machine learning and pattern recognition communities. In particular, the research field devoted to the study of adaptation algorithms aimed at overcoming the shift in probability distributions is referred to as Domain Adaptation and falls under the broader field of *transfer learning* [Pan and Yang, 2010][1]. Transfer learning is a more general sub-field of machine learning, which is itself closely related to *multi-task learning* [Thrun and Pratt, 1998]. Indeed, transfer learning is concerned with the development of solutions for problems involving not only different underlying distributions, but also different learning tasks, i. e. classification or regression scenarios in which both the output label space $\mathcal{Y}$ and the associated predictive function $f(\cdot)$ change.

4.2 NOTATION AND DEFINITIONS

Source and target domains

As in this Thesis we focus our attention to the changes in probability distributions, in the present Section we fix the notation and introduce the concepts necessary to tackle the adaptation problem in these situations. The definitions below build on the initial machine learning terms presented in Section 3.2. In general, the field of investigation of DA aims at leveraging the information collected in a given *source domain* $\mathcal{D}_S$ for its use in a different but related *target domain* $\mathcal{D}_T$. Throughout the Thesis, subscripts "$\cdot_S$" and "$\cdot_T$" will be used to denote elements related to the source and target domains, respectively. A domain $\mathcal{D}$ consists of some input variables X and associated output labels Y, governed by a joint probability distribution $P(X,Y)$.

Types of shift

Based on the observed change in the probability distributions, several distinct types of shift can be distinguished [Quiñonero-Candela et al., 2009, Moreno-Torres et al., 2012]. In the most general case, when the joint source and target distributions differ, i. e. $P_S(X,Y) \neq P_T(X,Y)$, the problem is referred to as *dataset shift*. A more specific situation called sample selection bias is encountered when a constraint affects the sampling process: $P_S(X,Y) = P(X,Y|\delta = 1)$ while $P_T(X,Y) = P(X,Y)$, where δ is a binary selection variable. This means that, even though the general distribution $P(X,Y)$ controlling the two domains is the same, in the source domain a bias in the selection of the samples occurs, i. e. its training set will only cover a portion of the support of the complete distribution. Other more technical terminologies include the distinction between *covariate shift*, where only the distribution of the input variables (covariates) changes ($P_S(X) \neq P_T(X)$), *prior probability shift*, where the prior distribution of the labels evolves due to an imbalance in the class counts ($P_S(Y) \neq P_T(Y)$) and *concept shift*, when only the class-conditional distributions or the posterior distributions of the classes change ($P_S(X) = P_T(X)$ but $P_S(X|Y) \neq P_T(X|Y)$ or $P_S(Y|X) \neq P_T(Y|X)$) [Moreno-Torres et al., 2012]. In what follows, we consider general dataset shift problems.

[1] Because both Domain Adaptation and transfer learning are very young research areas, a unifying terminology and common definitions are still lacking. Nonetheless, throughout this Thesis, the designation, description and usage of the different concepts is consistent with the definitions of Section 4.2, which, in turn, may differ from those found in specific papers.

Table 4.1: Categorization of standard and DA learning problems based on the availability of class labels in each domain.

$P(X_S, Y_S) =$ $P(X_T, Y_T)$?	Y_S available ?	Y_T available ?	Approach
✓	✗	✗	Standard unsupervised learn.
✓	✓	✓	Standard supervised learn.
✗	✓	✗	Unsupervised DA
✗	✓	✓	Supervised DA

Types of adaptation problems

Restricting our view to classification problems, a categorization of the different DA approaches with respect to the availability of labels in the source and target domains can be attempted as follows. Let $D_S = \{X_S, Y_S\} = \{(\boldsymbol{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$ be the set of n_S labeled source data and $D_T = \{X_T, Y_T\} = \{(\boldsymbol{x}_{T_j}, y_{T_j})\}_{j=1}^{n_T}$ the set of n_T labeled target data, with samples $\boldsymbol{x}_{S_i} \in \mathbb{R}^{d_S}\, \forall\, i$ and $\boldsymbol{x}_{T_j} \in \mathbb{R}^{d_T}\, \forall\, j$. The scope of all adaptation techniques is to predict the class labels $y_{T,\text{test}}$ for some unseen target test samples $\boldsymbol{x}_{T,\text{test}} \in \mathcal{D}_T$ based on the labeled information D_S in the source domain. In this context, strategies that are allowed to resort to labeled samples also in the target domain, i. e. the pairs $(\boldsymbol{x}_{T_j}, y_{T_j}) \in D_T$, are termed (fully) *supervised DA* approaches. Conversely, *unsupervised DA* methodologies predict target labels based exclusively on the use of labeled data from D_S in the training phase and/or for the definition of the adaptation strategy (no access to Y_T). Table 4.1 summarizes these different types of knowledge transfer across domains and relates them to classic same-domain learning problems. Regarding this Thesis, the research presented in Chapter 6 and 9 deals with a supervised DA exercise, whereas Chapters 7 and 8 fit into an unsupervised DA context.

General assumptions

It is generally assumed that the dimensionality of the two domains is the same and amounts to d, i. e. $d_S = d_T = d$. However, methodologies that are independent of the data dimensionality can be envisaged (see Chapter 9). Likewise, most of the methods are developed under the hypothesis of a common set of c classes, that is both y_{S_i} and y_{T_j} generally can only take the same c labels.

4.3 ASSESSING DISTANCES BETWEEN DISTRIBUTIONS

A necessary tool

In a first stage of an analysis involving more than one dataset it is crucial to quantify the importance (and the type) of the dataset shift occurred between source and target domains. Therefore, objective and robust measures of distance between distributions are needed. Hereafter, we present the main existing parametric measures along with a novel distribution-free kernel-based metric.

4.3.1 *Parametric distance measures*

Measures developed in statistics

In the literature, many measures have been used to evaluate the statistical difference between probability distributions: *Kullback-Leibler divergence* [Kullback and Leibler, 1951], *Jensen-Shannon divergence* [Lin, 1991], *Bhattacharyya distance* [Bhattacharyya, 1943], *Jeffries-Matusita* (JM) *distance* [Toussaint, 1972] are among the most popular. These distances have been developed and thoroughly employed in the field of statistics. More specifically, in the remote sensing community, the attention has focused mainly on the last two metrics of this list. These measures have mainly been used to estimate class-separability when comparing filter *feature selection* techniques [Serpico and Bruzzone, 2001] or to evaluate the invariance of the selected features over different spatial domains [Bruzzone and Persello, 2009].

Formulas

The distance between the distributions associated with an unlabeled source dataset X_S and another unlabeled target dataset X_T is provided by the above-mentioned distance measures as:

- Bhattacharyya distance

$$\mathrm{B}(X_S, X_T) = -\ln\left(\int_{\boldsymbol{x}} \sqrt{p(\boldsymbol{x}_S)p(\boldsymbol{x}_T)}\, \mathrm{d}\boldsymbol{x}\right) \;, \tag{4.1}$$

- JM distance

$$\mathrm{JM}(X_S, X_T) = \sqrt{\int_{\boldsymbol{x}} \left(\sqrt{p(\boldsymbol{x}_S)} - \sqrt{p(\boldsymbol{x}_T)}\right)^2 \mathrm{d}\boldsymbol{x}} \;. \tag{4.2}$$

The two quantities are intimately related, so that the latter can be computed from the former as

$$\mathrm{JM}(X_S, X_T) = \sqrt{2\Big(1 - \exp\big(-\mathrm{B}(X_S, X_T)\big)\Big)} \;. \tag{4.3}$$

Concretely, under the assumption that both X_S and X_T follow multivariate Gaussian distributions defined by mean vectors $\boldsymbol{\mu}_S$ and $\boldsymbol{\mu}_T$ and by covariance matrices Σ_S and Σ_T, the Bhattacharyya distance becomes

$$\begin{aligned}\mathrm{B}(X_S, X_T) =& \frac{1}{8}(\boldsymbol{\mu}_S - \boldsymbol{\mu}_T)^\top \left(\frac{\Sigma_S + \Sigma_T}{2}\right)^{-1} (\boldsymbol{\mu}_S - \boldsymbol{\mu}_T) \\ &+ \frac{1}{2}\ln\frac{|(\Sigma_S + \Sigma_T)/2|}{\sqrt{|\Sigma_S||\Sigma_T|}} \;. \end{aligned} \tag{4.4}$$

When investigating class discrimination, the JM distance is usually preferred to the Bhattacharyya distance since, for increasingly separable classes (mean vectors moving far apart), the JM distance will saturate at $\sqrt{2}$ when the classes do not overlap anymore (maximal classification accuracy, in Bayesian sense), while the Bhattacharyya distance will continue to grow.

4.3.2 *Maximum Mean Discrepancy*

Strengths of MMD

The previously presented distance measures can be affected by data dimensionality (they are based on the Mahalanobis distance) and by the presence of multimodal distributions (they assume unimodal Gaussian PDFs). To cope with these problems, we introduce a recently presented metric for comparing distributions, the *Maximum Mean Discrepancy* (MMD) [Borgwardt et al., 2006, Gretton et al., 2012]. MMD is based on the difference of the mean of the distributions computed in a common RKHS. This non-parametric kernel-based measure can be easily calculated also in presence of a large number of variables describing the data points. Furthermore, it is able to finely detect distribution shifts even when these appear under the form of additional modes in the distribution. MMD has previously been used to attribute different weights to shifted training and test samples when trying to match their distributions in the RKHS [Huang et al., 2007]. Likewise, in Gomez-Chova et al. [2010] the authors also exploit this mapping to evaluate cluster similarity by computing the difference of the means of sets of samples in the feature space.

The formulation

The empirical estimate of the MMD between the distribution of source data X_S and that of related target data X_T is given by

$$\mathrm{MMD}(X_S, X_T) = \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} \varphi(\boldsymbol{x}_{S_i}) - \frac{1}{n_T} \sum_{j=1}^{n_T} \varphi(\boldsymbol{x}_{T_j}) \right\|_{\mathcal{H}}^2 , \qquad (4.5)$$

where $||\cdot||_{\mathcal{H}}$ is the ℓ_2 norm computed in a RKHS induced by φ. Thus, MMD is the squared distance between sample means in this feature space and approaches zero when the two distributions tend to be exactly the same. Taking advantage of the kernel trick one can rewrite (4.5) as:

$$\mathrm{MMD}(X_S, X_T) = \left(\frac{1}{n_S^2} \sum_{i,j=1}^{n_S} K(\boldsymbol{x}_{S_i}, \boldsymbol{x}_{S_j}) - \frac{2}{n_S n_T} \sum_{i,j=1}^{n_S, n_T} K(\boldsymbol{x}_{S_i}, \boldsymbol{x}_{T_j}) + \frac{1}{n_T^2} \sum_{i,j=1}^{n_T} K(\boldsymbol{x}_{T_i}, \boldsymbol{x}_{T_j}) \right)^{1/2} = \mathrm{Tr}(\boldsymbol{K}\boldsymbol{L}) , \qquad (4.6)$$

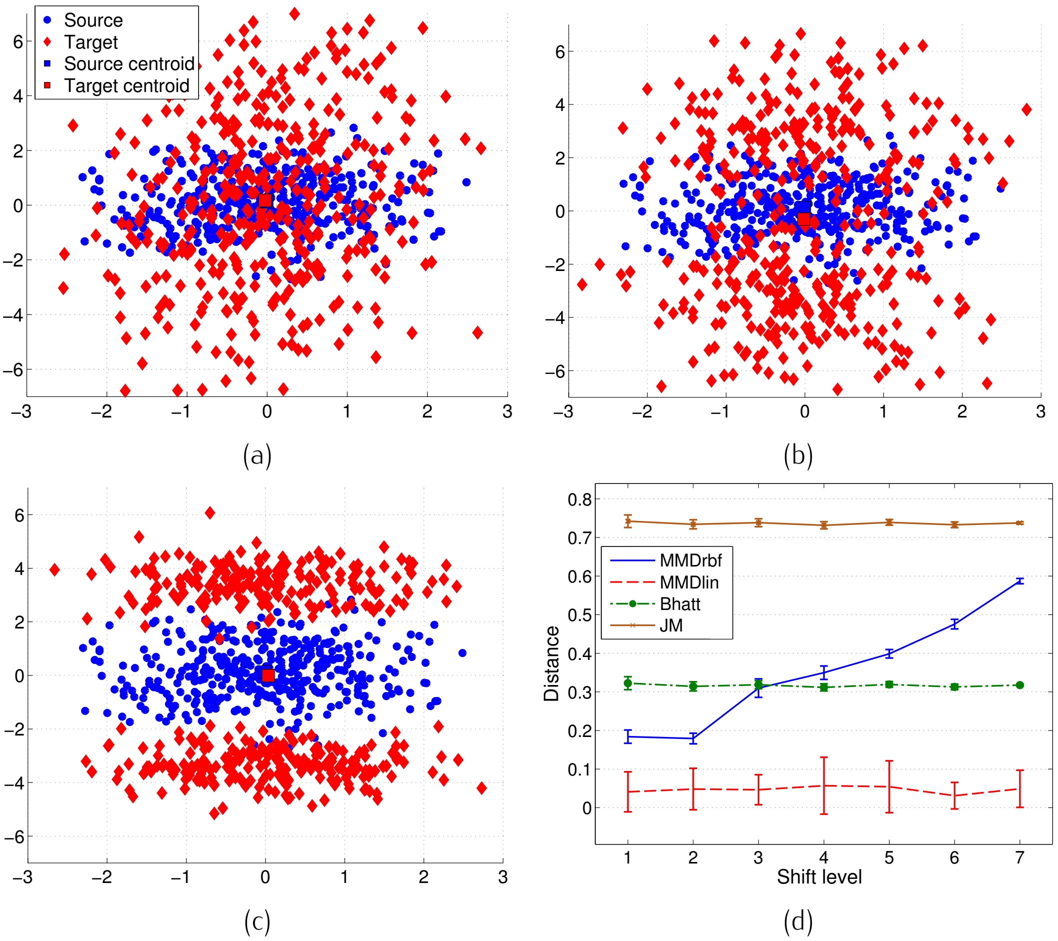
where

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_{S,S} & \boldsymbol{K}_{S,T} \\ \boldsymbol{K}_{T,S} & \boldsymbol{K}_{T,T} \end{pmatrix} \in \mathbb{R}^{(n_S+n_T)\times(n_S+n_T)} , \qquad (4.7)$$

with $\boldsymbol{K}_{S,S}, \boldsymbol{K}_{T,T}, \boldsymbol{K}_{S,T}, \boldsymbol{K}_{T,S}$ being the kernel matrices reflecting data similarities in the source domain, target domain and across domains, respectively. Matrix $\boldsymbol{L}$ contains the coefficients combining the elements of kernel matrix $\boldsymbol{K}$ to obtain (4.5). If $\boldsymbol{x}_i, \boldsymbol{x}_j \in X_S$: $L_{i,j} = 1/n_S^2$, if $\boldsymbol{x}_i, \boldsymbol{x}_j \in X_T$: $L_{i,j} = 1/n_T^2$, and otherwise: $L_{i,j} = -1/n_S n_T$.

Requirements

Theoretically, to detect subtle distribution differences, the kernel function $K(\cdot,\cdot)$ should be chosen as a universal kernel, e. g. Gaussian RBF or Laplace kernels [Gretton et al., 2012]. Practically, it can be any positive



(a)

(b)

(c)

(d)

Figure 4.1: Dataset shift assessment on a toy dataset consisting of source (**blue circles**) and target (**red diamonds**) domain combinations with increasing levels of dataset shift (one realization is shown). (a) shift level #1, (b) shift level #4, (c) shift level #7. (d) Distance between domains (average and standard deviations over 10 experiments) as measured by the MMD computed using a Gaussian RBF kernel with $\sigma = \sqrt{2}$ (`MMDrbf`), the MMD with a linear kernel (`MMDlin`), the Bhattacharyya distance under a unimodal Gaussian assumption (`Bhatt`), the JM distance under a unimodal Gaussian assumption (`JM`).

semi-definite function [Jegelka et al., 2009], hence fulfilling Mercer's conditions (see Section 3.3.3). In the linear case $K_{i,j} = \mathbf{x}_i^\top \mathbf{x}_j$, the MMD measure reduces to be the simple difference of the means of the two distributions in the input space. Thus, for such an indicator the use of an appropriate non-linear kernel function is key.

4.3.3 *A toy example*

Experimental setup

To illustrate the importance and the effectiveness of the non-linear mapping induced by the kernel trick for MMD, we resort to Fig. 4.1. In panels (a), (b) and (c), this figure presents a toy dataset consisting of source domain data being kept fixed and target domain data displaying an increasing level of shift with respect to the source domain. The source distribution is a unimodal bi-variate Gaussian distribution. Initially, the target distribution is also a unimodal Gaussian but throughout 7 shift levels, it turns into a mixture of 2 Gaussians migrating farther apart along the vertical axis. The variance of these 2 components is gradually reduced to keep a constant common covariance matrix. Note that both distributions are always centered at zero mean, i. e. $\boldsymbol{\mu}_S = \boldsymbol{\mu}_T = (0, 0)$. Shift level #1 (Fig. 4.1(a)) represents the initial situation where the target dataset presents a single mode but the point cloud has a higher variance with respect to the source data. Reaching shift level #7 (Fig. 4.1(c)), the target distribution is made up of two Gaussians centered at $(0, 3.4)$ and $(0, -3.4)$, respectively. We considered 10 random realizations of the datasets consisting of 400 samples per domain and measured the distance between them with the previously presented metrics.

Results

By examining the results presented in Fig. 4.1(d), the effectiveness of the MMD indicator in detecting the distribution shift appears as striking. Indeed, the MMD using a Gaussian RBF kernel steadily grows as the shift level increases. The ability of the MMD to properly capture the progressive evolution in the shape of the data is guaranteed by the non-linear mapping ensured by the chosen kernel function. This is highly encouraging since the first two moments of the overall target distribution (mean vector and covariance matrix) remain unchanged throughout the 7 stages. Contrarily, the MMD in its linear version (with a linear kernel) does not detect the dataset shift at all because it simply measures the distance between the means in the input space. This is exactly the distance between centroids represented with squares in the plots, which remain very close, yielding a distance value close to 0. The parametric measures also fail in highlighting the change in probability distributions, returning constant distance values across the entire range of shift levels. In this case, the assumption on the underlying distribution needed by both the Bhattacharyya and JM distances is what precludes the correct assessment of the shift. Indeed, without a prior knowledge on the PDF generating the data (the case generally encountered in real-life problems), these metrics assume a Gaussian distribution with a single mode, an over-simplification of the actual situation. Nonetheless, as the covariance of the source and target domains is different, we draw the attention to the fact that the distance values are at least far above zero.

4.4 FAMILIES OF ADAPTATION APPROACHES

Categorization based on "what" and "how" to adapt

The approaches to DA that have been recently proposed in the literature can be divided into three main categories [Pan and Yang, 2010, Margolis, 2011]. Focusing on the problem of classification, such a grouping is based on what type of knowledge is being transferred across domains and how this transfer takes place.

1. The first category of strategies concerns *instance-transfer* approaches, where samples of the source domain are reweighted for their further

use in the target domain or taken as initial training set to define active queries on this newly acquired data. In Chapter 6, we propose a method belonging to this family.

2. The second category regards the *feature-representation-transfer* framework. The purpose of this type of techniques is to change the input space of our datasets to obtain a new set of shared and invariant features. The methods utilized here are rooted in the machine learning sub-fields of FE, feature selection, *manifold learning*, etc. The differences in the statistical distribution between the two domains should thus be maximally reduced, resulting in an improved portability of the classifiers across domains. Chapters 7, 8 and 9 will deal with this type of adaptation strategy.

3. The third category is related to methods aiming at transferring and adapting the parameters of a classification model to be applied in the target domain. These solutions directly adapting the classifier itself are termed *parameter-transfer* approaches.

4.4.1 *Instance-transfer*

Instance reweighting

Among the approaches relying on a sample-based transfer, one of the main directions of research is that of instance reweighting techniques. In Dai et al. [2007], the authors propose *TrAdaBoost*, an adjustment to DA of the *AdaBoost* method [Schapire, 1999]. As it will be explained in more details in Chapter 6, TrAdaBoost is a technique designed to iteratively change the weight of the data points in each domain. The goal is to increase the attention of the classifier to incorrectly predicted samples in the domain of interest, the target domain, while reducing the impact on the final solution of misclassified source samples. Along this line of thought, Eaton and desJardins [2009] extend the approach to handle multiple source domains. They introduce an additional reweighting factor accounting for the quality of the knowledge transfer provided by each of the source domains. This attempts to prevent the situation referred to as *negative transfer*, encountered when the information transferred from the source domain actually hampers the prediction task in the target domain.

Importance sampling

Another ensemble of techniques falling in the instance-transfer category is constituted by *importance sampling* approaches. Under the hypothesis of simple covariate shift, where only the marginal probabilities are assumed to have changed ($P_S(X) \neq P_T(X)$), the key insight behind all of these techniques is the following. Each source sample $\boldsymbol{x}_{S_i}$ carries an importance for the target domain prediction that can be appropriately determined by estimating $P_T(\boldsymbol{x}_{S_i}) / P_S(\boldsymbol{x}_{S_i})$. This ratio of the target to source domain probability is then used as a weight to raise the impact of those labeled source samples lying in a region with a high density of target samples. Since in practice $P_S(X)$ and $P_T(X)$ are often unknown, the interest has focused on approaches that do not require an explicit modeling of such probabilities. The previously cited ratio is instead directly estimated. In this context, Huang et al. [2007] introduce a technique based on MMD called *Kernel Mean Matching* (KMM) to retrieve the ratio by matching, via reweighting, the means of the source and target domains in a kernel-induced feature space. In parallel to this work and with the same goal, Sugiyama et al. [2007] propose an algorithm which minimizes the Kullback-Leibler divergence from the true target probabilities to their estimate produced via a rescaling of the source probabilities.

Adaptive AL strategies

A third sub-category we term *adaptive Active Learning* concerns the approaches resorting to AL techniques to smartly sample the new domain. Such a group can be considered part of the instance-transfer category because the samples of the source domain are somehow re-used as the starting point for the sampling of the target domain. Recent advances in machine learning show that the combination of the AL and DA frameworks is effective. In Shi et al. [2008], a principle apt to reduce the number of examples to be labeled by the user is outlined. The authors suggest to use the classifier trained in the source domain to obtain a prediction for the relevant target instances proposed by an AL strategy. At this point, if the confidence on the prediction is too low, an expert is asked to provide the correct label for the sample. Later on, Rai et al. [2010] proposed a preprocessing step highlighting the interesting regions of the target domain in order to reduce the size of the set of candidate samples for the AL search. Based on this contribution, the same authors outline a complete framework for AL in a DA setting [Saha et al., 2011]. In Chapter 6, we will consider a combination of instance reweighting and AL to achieve adaptation.

4.4.2 *Feature-representation-transfer*

The principle

The purpose of feature-representation-transfer strategies is to find a common representation of the source and target datasets that minimizes the differences between these two domains while maintaining their main data properties. Once the samples are mapped to the same space defined by the new features (either with a projection to a common subspace or with a cross-domain conversion of the input spaces), a classifier is trained in the source domain using the available labeled examples, and then inference is performed directly in the aligned target domain.

Feature extraction

A first sub-category of these feature-representation-transfer approaches is constituted by FE techniques. In Pan et al. [2008], the researchers have been concerned with the development of *Maximum Mean Discrepancy Embedding*, a technique that aims at minimizing the MMD between the projections of the source and target domains. After solving a kernel learning problem, the technique extracts a new set of common features for the two domains embedding the samples in a shared low-dimensional sub-space. However, the procedure has many drawbacks, the most important of which is that it can not generalize to unseen samples. Subsequently, the same authors extended these preliminary findings to develop *Transfer Component*

Analysis (TCA) [Pan et al., 2011], an out-of-sample and faster formulation of the previous method. More details about this approach and, namely, a thorough study of unsupervised TCA and its semisupervised extension named *Semisupervised Transfer Component Analysis* (SSTCA), will be provided in Chapter 7.

Manifold alignment

Another line of research is represented by *manifold alignment* [Ham et al., 2003]. As for FE strategies, the purpose is find a new latent space minimizing the distance between the data points of the same class coming from different domains. The key property of these approaches considering the manifold is that they try to maximally preserve the original local structure of the data. In this context, a suite of effective approaches has been proposed by Wang and Mahadevan [2008, 2009, 2011]. All these works encode the local similarities through a *graph Laplacian* based on a nearest neighbor adjacency matrix. The proposed methods are designed to handle datasets presenting a different dimensionality, a characteristic that makes them highly flexible. In Gopalan et al. [2011], Gong et al. [2012], Gopalan et al. [2013], a slightly different perspective is adopted. The authors assume the existence of a series of intermediate subspaces gradually connecting the source and target domain manifolds. They propose to compute the projections of the data on these subspaces and train a classifier therein.

Feature selection

Instead of extracting features anew, Chen et al. [2011a] propose an approach centered on *feature selection* that builds on a previous work by Satpal and Sarawagi [2007]. In an iterative procedure extending the training set both with new samples and with new features, they try to promote the usage of those features that behave similarly in both domains, namely in the source training set and in the target test set.

Augmented input space

An important ensemble of techniques devised especially in the field of *natural language processing*, is constituted by strategies augmenting the initial input space with new variables. The stacked set of features is then utilized to build a transfer model. In Daumé III [2007] and in Daumé III et al. [2010], the authors propose a kernel-based feature replication approach with a semisupervised extension that proved potential in sequence labeling tasks. Other relevant works concern the development of algorithms such as *Structural Correspondence Learning* [Blitzer et al., 2006, 2007] and *Spectral Feature Alignment* [Pan et al., 2010]. These techniques provide appropriate solutions to problems such as sentiment classification for opinion mining [Pang et al., 2002]. In this type of application where a human expert is asked to decide the polarity of a product review, much interest has been devoted to algorithms able to suitably transfer the classification rules from one product domain to another (for instance from the digital cameras to the video games domain) [Blitzer et al., 2007]. These methodologies exploit the fact that, working with *bag-of-words* models, in the two types of datasets both *domain-independent* and *domain-specific* features will naturally appear. The former (e. g. terms such as "good", "bad" or "never buy") are considered pivot features allowing to bridge the gap between domains, whereas the latter (e. g. terms such as "compact" and "blurry" for digital cameras or "realistic" and "boring" for video games) are the features to be aligned by finding the proper cross-domain correspondences.

Dictionary Learning

A last sub-group of techniques tackling adaptation through changes in the representation of the features consists of DL-based methods. In Ni et al. [2013] a procedure to gradually capture the dataset shift is outlined. The source and target domains are represented with their respective dictionaries and a virtual path easing the knowledge transfer among them is defined thanks to a series of intermediate dictionaries connecting the two ends. Shekhar et al. [2013], instead propose an algorithm that first embeds the samples of the two domains into a low-dimensional common sub-space and then learns a shared dictionary suitably representing both. Subsequently, a classification based on the reconstruction error is carried out. In Wang et al. [2012], the authors propose a cross-domain image synthesis approach to directly convert one into another the input spaces of images of different styles (e. g. sketch vs. photo) by means of a linear transformation. The algorithm simultaneously learns a dictionary pair (one per image) and a mapping function from one to the other. A more detailed explanation on this work will be provided in Chapter 9.

Unsupervised vs. supervised

Within this category of feature-representation-transfer strategies we can make three further distinctions as regards the basis for the alignment. First, when the samples from both domains used to define the change in the input space are all unlabeled, such transformation is unsupervised. In general, a joint-domain FE via PCA, KPCA and TCA (see Chapter 7), or the HM procedure in remote sensing (see Chapters 7 and 8) are all good examples of such a category of techniques. Second, if the employed method makes use of the available labeled data in the source domain, it can be defined as a supervised transformation (see SSTCA in Chapter 7). Third, when also target labeled data come into play in the definition of the alignment, the procedure modifying the feature representation could be termed fully supervised (see the cross-image synthesis based on DL of Chapter 9). However, note that this grouping only concerns the type of change of the initial input spaces prior to the actual cross-domain classification. The entire methodology, with its alignment and classification phases, will still be fitting under either the unsupervised DA or the supervised DA category.

4.4.3 *Parameter-transfer*

SVM-based techniques

Lastly, we review a perhaps less developed family of adaptation strategies, the parameter-transfer approaches. These methods deal with the adjustment of the classifier itself and have often been adopted in combination with SVM classifiers. In Yang et al. [2007a] the researchers propose an *Adaptive SVM* for video classification to adjust several SVMs initially trained on multiple source domains (called auxiliary datasets). The goal is to learn a specific delta function to be added the original decision function in order to properly model the instances of the new target domain. An AL extension of the latter method has been then proposed in Yang et al. [2007b] by the same authors. Within this family of methodologies as well, the ability of MMD in detecting the distribution mismatch has been used as foundation to develop new DA algorithms. Duan et al. [2009] suggest to combine the structural risk functional of the SVM and the MMD in a joint minimization problem. By further generalizing the approach, in Duan et al. [2012] the authors develop a combination of the MMD minimization principle with a multiple-kernel learning framework. With a different perspective, Bruzzone and Marconcini [2010] propose to deform the SVM classifier by discarding contradictory old source training samples with respect to the distribution observed in the target domain.

Part III

DOMAIN ADAPTATION APPROACHES FOR REMOTE SENSING IMAGE CLASSIFICATION

5 REVIEW OF DOMAIN ADAPTATION STUDIES IN REMOTE SENSING

Outline: *This Chapter acts as a bridge between the preceding Part ii of this manuscript, describing the machine learning and Domain Adaptation frameworks, and the next Chapters 6, 7, 8 and 9 where, from that standpoint, we seek a solution to remote sensing problems. In Section 5.1, the challenges encountered in the field of Earth observation are translated into Domain Adaptation terms. The remote sensing concepts are linked with the statistical notation that will be adopted in the following. In Section 5.2 we include an overview of the early approaches developed to extend classification systems beyond single images. Subsequently, starting from the same three previously defined categories of Domain Adaptation techniques, we provide an exhaustive review of the current state-of-the-art of the adaptation approaches in remote sensing.*

5.1 THE CONTEXT

Recap

As mentioned in Section 2.5.2, when the atmospheric and ground conditions or the geometry of the acquisition vary from one image to another, a series of physical phenomena (list available on page 22) induces a shift in the probability distribution of the spectra of the different land-cover classes. Another relevant problem concerns incomplete reference data. In many applications, the user is interested in classes that are far from being pure and uniform over the scene. When the sampling areas are small and very localized, this can lead to a bias in the composition of the training set, with reference data only partially covering the complete class distribution. As a consequence of these two types of issues, we observe a shift in the statistical distributions of the pixels and the land-cover classes. This compromises a direct knowledge transfer from one image to another or among portions of a large scene. Since land-cover models are dependent on the spectra observed under specific acquisition conditions, they tend to generalize poorly when applied to new environments. Under such constraints, the development of large-scale VHR land-cover/land-use mapping systems that require multiple or extended remote sensing images is hindered.

Link between machine learning and remote sensing

As we discussed in Section 2.5.1, Earth observation scientists have recently started tackling the above-mentioned model portability problems from the standpoint of statistical learning (see Section 3.2). In this context, each image possessing an associated ground truth can be represented by a set of d spectral bands X with the thematic class labels Y. Using a matrix notation, a set of n labeled pixels is denoted by $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with $y \in \mathcal{C} = \{1, 2, \ldots, c\}$, the set of c land-cover classes.

Link between DA and remote sensing

Looking more specifically at DA applied to remote sensing image analysis, following the notation introduced in Section 4.2, we can state the underlying problem in this fashion. We consider each image acquisition a separate domain. The goal of portability is reached when a classifier trained on a given source image $\mathcal{D}_S$ governed by $P_S(X, Y)$ is successfully applied on a target image of interest $\mathcal{D}_T$ whose pixels are assumed to be drawn from a slightly different but related probability distribution $P_T(X, Y)$. The same applies to models developed on portions ($\mathcal{D}_S$) of large images and applied to predict the land-cover over their entire extent ($\mathcal{D}_T$). We name the former a case of dataset shift ($P_S(X, Y) \neq P_T(X, Y)$) while we refer to the latter as an instance of sample selection bias ($P_S(X, Y) = P(X, Y|\delta = 1)$ and $P_T(X, Y) = P(X, Y)$), the selection variable δ being the spatial constraint preventing the sampling of the training set from the exhaustive distribution $P(X, Y)$ governing the whole image. As mentioned earlier, for simplicity, it is assumed that the two domains however bear some key similarities: they should share the same set of d spectral bands (same sensor having acquired the images), i. e. $d_S = d_T = d$, and c land-cover classes (no new classes should appear in the target image), i. e. $c_S = c_T = c$. Still, in the next Section 5.2 and in Chapter 9 we will see some exceptions to these rules.

The challenges

By adapting and applying the techniques developed in the field of DA we reviewed in the preceding Chapter 4, or simply by taking inspiration from the novel concepts proposed, the remote sensing community has taken a notable step forward. The common goal of DA strategies in remote sensing thematic classification tasks is to be able to accurately predict the land-cover in the target image. The procedure is based on an initial training set $\{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$ composed of usually abundant ground truth data existing on a given source image. Starting from the types of learning problems listed in Tab. 4.1 on page 49, in Earth monitoring applications we can draw a distinction between two main settings in which DA can take place.

- *Supervised DA remote sensing problems*: the user has access to a small set of labeled samples $\{(\mathbf{x}_{T_j}, y_{T_j})\}_{j=1}^{n_T}$ in the target image. This implies that the model initially built based on source data can be refined for the prediction in the target domain. The assumption of a large ground truth for the new images is both unrealistic and void of any interest. In fact, in most applications $n_T \ll n_S$ will hold, as the target image is generally a very recently acquired image not yet analyzed nor sampled. Moreover, if $n_T > n_S$, models built exclusively based on target samples would outperform those based on the less thoroughly labeled and shifted source domain. There would be no need to resort to previously labeled acquisitions.

- *Unsupervised DA remote sensing problems*: the user has no access to any sample in the target image. This is clearly a more challenging yet common situation. The newly acquired target images have to be classified relying solely on models developed on labeled pixels of the source image. Such an approach is preferred when a large number of new images is collected and a rapid batch processing of the series of acquisitions is required. In this case, after that suitable adaptation measures are adopted, an antecedently developed source classifier can be directly applied to predict the land-cover on the target images (i. e. without re-training).

5.2 LITERATURE REVIEW

In this Section we will review the existing state-of-the-art DA approaches to remote sensing model portability problems. As already mentioned, this is a relatively new research direction, thus mostly encompassing literature from the last decade concentrating on the study of VHR and hyperspectral imagery. For the sake of coherence, we will continue with the distinction of the approaches in the same three categories previously identified in Section 4.4. Before that, in the next Section, we will focus on a brief review of the first pioneering moderate resolution approaches using multispectral acquisitions.

5.2.1 *Signature extension approaches*

The beginning

When pursuing thematic image classification at a large scale, the portability of land-cover classifiers across acquisitions has been initially studied in the signature extension framework. This field of investigation has been a lively research area ever since the beginning of the Landsat mission [Fleming et al., 1975], mainly focusing on moderate resolution applications.

Forestry

Among the early works, in Pax-Lenney et al. [2001], the researchers have evaluated the extension of the predictions of a neural network classifier across space and time on Landsat TM imagery for forestry applications. In doing so, they gauged the efficiency of simple atmospheric compensation methods with respect to more sophisticated physically-based approaches. Additionally, attention was paid to the influence of the seasonal effects. In Woodcock et al. [2001], the authors proceeded with a cross-sensor approach to combine Landsat 5 TM and Landsat 7 ETM+ acquisitions to ease binary forest change mapping efforts over nearby scenes. Foody et al. [2003] investigated the transferability across validation sites located in different tropical regions of the world of predictions of forest biomass based on Landsat TM data. In this context, they assessed the performances of vegetation indices, multivariate regression and neural networks. Nonetheless, they witnessed poor portability results in general.

Land-cover mapping

In Olthof et al. [2005], a study of the extension of northern land-cover spectral signatures is proposed. By comparing it to an absolute atmospheric compensation method, the authors observed superior classification accuracies for a relative normalization method based on a calibration with respect to low resolution SPOT imagery. Moreover, due to the change in vegetation composition, the authors noted a marked decrease in accuracy when the knowledge transfer took place in the north-south direction over large latitudinal extents. Finally, Knorn et al. [2009] exploited the overlap existing between neighboring Landsat images to set up a chain classification procedure based on SVMs.

5.2.2 *Instance-transfer approaches in remote sensing*

Classic AL

As regards DA, the remote sensing community has mostly applied instance-transfer techniques via AL procedures to iteratively sample the new image while initially relying on source samples. General purpose AL techniques have been widely studied in the remote sensing community during the last years [Rajan et al., 2008, Tuia et al., 2009b, 2011b, Demir et al., 2011, Volpi et al., 2012c, Di and Crawford, 2012]. Indeed, when analyzing a given scene, procedures allowing the user to optimally select the pixels to label can dramatically reduce the sampling burden. Smartly built training sets also yield classification models more effectively discriminating the land-cover classes [Crawford et al., 2013].

Adaptive AL strategies

The application of these principles in a cross-domain setting is enticing: a classifier trained on a first acquisition can be adapted to a new image with a minimal effort by finding the pixels representing the shift between the two images. Under this perspective, adaptive AL can be considered a fully supervised DA approach. Such a principle was firstly explored by Jun and Ghosh [2008]. The authors showed that the proper adaptation could be achieved by actively querying, pixel by pixel, the target samples necessary to be integrated in the knowledge transfer process. They combined these active queries with a reweighting concept similar to that proposed in Dai et al. [2007]. The AL method they employed is based on the Kullback-Leibler divergence and is thus constrained by data normality assumptions. In Tuia et al. [2011a], AL has been proposed for the correction of sample selection bias when dealing with a training set issued from a small sub-region of an image. In this setting, they studied the ability of an adaptive system based on pre-clustering to discover previously unknown classes appearing in the target domain (the complete image). In this case it is thus assumed that the target domain possesses more land-cover classes than the source domain, i. e. $c_T > c_S$. Successively, other methods specifically designed to re-use already collected source ground truth information to initialize the AL loop have been advised in Persello and Bruzzone [2012]. The authors propose to gradually remove along the iterations the source samples conflicting with the distribution of the classes in the target domain. A convergence criterion to know when to stop the iterative process without resorting to a test set is also put forward. Lastly, Alajlan et al. [2014] investigate the benefits of adaptive AL for classification problems at the continental scale by employing low spatial resolution MODIS data.

5.2.3 *Feature-representation-transfer approaches in remote sensing*

Classic approaches with single images

The second type of techniques, based on the feature-representation-transfer framework, constitutes also a relatively new research direction within the remote sensing community. However, few papers have addressed the analysis of multiple images through dimensionality reduction, despite the fact that the change of the *data space* (i. e. the input space in a machine learning sense) is a largely studied topic for single images. Considering FE, many different methods have been applied to single images to provide the end-user with either noise-free or more class-discriminant features [Arenas-García and Petersen, 2009, Kuo and Landgrebe, 2004, Li et al., 2011]. The same can be said for feature selection algorithms [Bruzzone et al., 1995, Tuia et al., 2010, Camps-Valls et al., 2010] aiming at subsetting the input space while preserving the physical meaning of the variables, with a significant number of contributions dating back to the 1970s [Narendra and Fukunaga, 1977]. Another line of research is represented by the strategies known under the name of *manifold learning* approaches. Those models have also been widely investigated for the application on single images [Bachmann et al., 2005, Yang and Crawford, 2012, Lunga et al., 2014]. Such approaches, particularly suitable for the analysis of hyperspectral images, rely on the local properties of the data to preserve their topology after the dimensionality reduction step.

FE with multiple images

When dealing with multiple images, the family of approaches based on FE generally comprises contributions that are focused on change detection applications. Nielsen et al. [1998] introduce the *Multivariate Alteration Detection* (MAD) technique to detect changes in bi-temporal images. Making use of the standard *canonical correlation analysis* [Hotelling, 1936], this FE method is aimed at finding suitable separate linear combinations of the initial spectral bands of the two images. Ideally, the difference between these newly extracted components (canonical variates) should bear a maximum of variance. The samples are projected into a space where the extracted components from the two images display similar values for the unchanged regions while maximally differing on the changed areas. The MAD framework has been extended with the inclusion of boosting-like procedures to iteratively increase weights for no-change pixels, and regularization to avoid singular covariance matrices, giving rise to *Iteratively Reweighted-Multivariate Alteration Detection* (IR-MAD) [Nielsen, 2007]. Moreover, preliminary results obtained with a kernel-based version of the technique are presented in Nielsen and Vestergaard [2013]. The MAD and IR-MAD transforms have also proved to be effective in detecting invariant regions in image time series to be used for relative radiometric calibration via regression analyses [Canty et al., 2004, Canty and Nielsen, 2008]. Starting from the same canonical correlation analysis, Volpi et al. [2013] present a semisupervised kernel-based FE method integrating both knowledge on unlabeled samples and a manifold regularization. Finally, the spectral alignment of bi-temporal images via FE has also been carried out with non-linear KPCA-based strategies as well [Nielsen and Canty, 2008, Volpi et al., 2012a]. It is worth noting that all the above methodologies are restricted to the study of spatially co-registered images, preventing thus their use in a standard cross-scene DA case.

Feature selection with multiple images

A feature selection approach to improve the generalization abilities of a classifier applied across disjoint portions of a hyperspectral image has been presented in Bruzzone and Persello [2009]. The authors pursue the selection of spectral channels that exhibit both a high spatial invariance throughout the image and a good class discrimination capability. To this end, a criterion function combining a *discrimination term* and an *invariance term* has been proposed to direct the search strategy for the identification of the best set of features.

Manifold learning with multiple images

The techniques exploiting the data manifolds for DA purposes can be considered as a means to further generalize standard semisupervised techniques [Shahshahani and Landgrebe, 1994] to the case where the unlabeled samples employed by the model belong to a dataset, the target domain, that follows a different probability distribution. In this framework, Kim et al. [2008] develop an iterative methodology to adapt a general land-cover classifier trained over a large area to adjust it for the prediction on a small, localized area. The technique, based on a regularization via the graph Laplacian, exploits local unlabeled samples appropriately reflecting the data distribution in the sub-region of the image where the prediction takes place. In Kim and Crawford [2010], the authors provide a complete framework for the application of manifold regularization to adapt the land-cover classification models. Under the assumption that no labeled samples can be obtained in the target domain, the knowledge transfer is suitably carried out among spatially disjoint areas of the same hyperspectral scene. Following a slightly different research direction, a manifold alignment approach is studied in Yang and Crawford [2011]. The authors address the problem of matching two datasets by seeking a joint manifold which incorporates prior features, i. e. the manifold of the source domain, and preserves the smoothness of the resulting aligned manifold. In Tuia et al. [2013a], after the application of a vector quantization algorithm to retrieve relevant centroids, adaptation is achieved by matching the shape of graphs defined thereon. These representations of the underlying data structure of the images are locally deformed and aligned to each other. Such a transformation of the manifolds is completely unsupervised and therefore is applicable in both directions, meaning that source and target images can interchangeably be taken as reference to adapt the other image. Jacobs et al. [2013] further refine this process to overcome the difficulties in handling large changes in the manifold structure and sub-optimal graph representations. As to the latter, they enhance the representation of the internal structure of the graphs for both domains by modeling them as two instances of a common underlying *Hidden Markov Random Field*. Finally, in Tuia et al. [2013b, In press.] the manifold alignment strategy of Wang and Mahadevan [2011] is smartly applied to map the images into a latent common space by means of two different projection functions. The latter are invertible and can even be defined for datasets possessing a different dimensionality, i. e. $d_S \neq d_T$. Hence, images taken from different sensors can be converted to the data space of another related image. This ultimately opens the opportunity for cross-sensor model portability. Although highly promising, the proposed methodology only works in supervised DA settings needing labeled samples from both images.

Summary

In general, one can readily see the great value brought to the field of remote sensing image classification by these feature-representation-transfer strategies. In fact, in most of the cases, the model for the target classification is exclusively built using labeled examples from the source image that have already been acquired. Thus, this family of methods is particularly suited for unsupervised DA tasks.

5.2.4 *Parameter-transfer approaches in remote sensing*

Pioneering DA works

This last category focusing on the models themselves, comprises the early DA papers by Bruzzone and Fernàndez Prieto [2001] and Bruzzone and Cossu [2002]. The common trait of both contributions is the type of application: the update of land-cover maps. In fact, the source domain is considered in this case to be the first acquisition of a multitemporal collection, while the target domain is represented by the images of the same area acquired at later times. The goal is to update the thematic classification map as soon as a new image enters the system but without requiring additional ground truth information. The authors term this exercise involving co-registered images a *partially unsupervised classification* task, a particular instance of unsupervised DA. In Bruzzone and Fernàndez Prieto [2001], pixels of the target domain are used to re-estimate parameters of the maximum likelihood classifier initially trained on the source image. The problem is solved by estimating, via *expectation-maximization*, a mixed density distribution (with as many components as the common number of classes) for the pixels of the target image. Building on the findings above, in Bruzzone and Cossu [2002], the researchers developed a cascade-classification approach to leverage the temporal correlation naturally observed between images of the same scene acquired at different time instants. In Bahirat et al. [2012], the Bayesian framework of the previous studies listed above is extended to handle a difference in the sets of land-cover classes observed in the multitemporal images.

Latest developements

In Rajan et al. [2006], ensembles of binary hierarchical classifiers are used to adapt to the target domain. Diversity in the predictions of the ensemble on the target image is used to reduce the number of binary classifiers to be combined via majority voting. Later on, Bruzzone and Marconcini [2009], propose to deform a SVM classifier by discarding old training samples that are contradictory with respect to the distribution observed in the target domain. At the same time, semi-labeled target samples are added to the training set. These samples are pixels of the target image whose tentative class labels are assigned by the adaptive classifier itself [Jackson and Landgrebe, 2001]. Additionally, the authors present an innovative circular accuracy assessment strategy for classifiers applied in a DA context, when no ground-truth is assumed available on the image of interest. In Gomez-Chova et al. [2010], knowledge transfer from source to target images is performed by matching the means of data clusters in a kernel-induced feature space with a procedure based on the same principles of MMD [Borgwardt et al., 2006] and KMM [Huang et al., 2007]. In Jun and Ghosh [2011], the authors use spatial detrending with a Gaussian process regression to compensate for spectral shifts that may have occurred in distinct regions of the image. Next, in Jun and Ghosh [2013], the latter approach addressing the spatial variations of the spectral signatures is extended to discover previously unknown land-cover classes. In Sun et al. [2013], the adaptation approach based on MMD and multiple-kernels of Duan et al. [2012] is applied to hyperspectral remote sensing data. Finally, Leiva-Murillo et al. [2013] apply to remote sensing the concepts of multi-task learning, a widely studied topic in machine learning. Working with SVMs, by sharing information across different classification tasks, they show improvements over approaches independently considering one task at a time.

6

SVM-BASED ADAPTIVE ACTIVE LEARNING VIA SAMPLE REWEIGHTING

Outline: *This Chapter is devoted to the study of adaptive Active Learning strategies. We will present an approach based on the TrAdaBoost algorithm to adequately reweight the samples of the source and target domain. The method takes advantage of the opportunities offered by Support Vector Machines in regard to weighting the training samples and returning probabilistic outputs. After the introductory Section 6.1 setting the context for the approach, the proposed iterative methodology consisting of two nested loops is outlined in Section 6.2. Next, Section 6.3 describes the datasets used and the setup of the experiments, while Section 6.4 reports and discusses the results. Finally, Section 6.5 summarizes the main achievements of this Chapter.*

6.1 INTRODUCTION

Combining AL and DA

Many DA strategies we reviewed in Section 5.2 assume that the labeled examples from the target image, when available, are passively obtained at once. However, if little resources can be allocated to the sampling and labeling of a given amount of new pixels, such sampling must be handled with care, in order to get maximal information from the limited number of queries. In this sense, the combined use of AL and DA approaches can be a winning strategy, since AL can be used to sample where source and target distributions differ. The DA component of such a hybrid system will make sure that the labeling effort could be further reduced by re-utilizing already collected ground truth associated with images acquired by the same sensor in a region with comparable characteristics.

Overview of the Chapter

In this Chapter, we propose to effectively combine the DA and AL frameworks in the context of SVM classification. The most informative pixels are sampled with active queries from the target image while adapting the obtained classifier using a transfer learning strategy, TrAdaBoost [Dai et al., 2007] (see Section 4.4.1), to leverage the original source data. The procedure fits in the framework we named adaptive AL. On the one hand, as base AL heuristic we apply the *Breaking Ties* (BT) strategy [Luo et al., 2005]. BT uses posterior class probabilities to rank the potential new training

The findings of this Chapter have been published in:

> G. Matasci, D. Tuia, and M. Kanevski. SVM-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):1335–1343, 2012.

samples according to their uncertainty for the current model. Note that, in any case, the AL procedure can be run using the sample selection heuristic that best suits the needs of the user (e. g. margin sampling or others, for a review see [Tuia et al., 2011b]). On the other hand, TrAdaBoost promotes a reweighting of the training instances provided to the classifier in order to assign a broader impact to key target domain samples while decreasing the influence of misleading source samples. This last step boosts the performance of traditional AL techniques when asked to intelligently suggest a sampling scheme in a target image whose class distributions have shifted. We propose an analysis of the performance of this procedure when combined with a SVM classifier accepting, in the optimization phase, weights associated with the training data samples. However, note that any supervised model allowing sample weights, such as the LDA classifier presented in Section 3.4, could be used.

Contributions

The purpose is to build a classifier that is able to efficiently handle the samples coming from the new image in order to provide a more accurate and adapted AL criterion. We provide a thorough illustration of the TrAdaBoost algorithm and an analysis of its behavior. Concerning the SVM classifier integrating weights for the instances, we study the separate evolution, with respect to the domain of membership, of the number of SVs and their weights during the AL procedure. Additionally, we carried out experiments studying the individual impact of the two approaches combined here: the active queries and the reweighting of the samples.

Main results

From the results, we can appreciate how both approaches are complementary and perform differently depending on the degree and complexity of the shift. Still, in all experiments, their combination resulted in an improved solution always providing the best accuracies. The sampling strategies are tested on two datasets. The first one concerns two QuickBird images of urban scenes while the second one implies a hyperspectral AVIRIS image of a natural environment. In both cases, experimental results prove the efficacy of the technique with respect to traditional non-adaptive AL approaches.

6.2 ADAPTIVE ACTIVE LEARNING

6.2.1 *SVM using instance weights*

Hereafter, we introduce the instance weighting SVM, the base classifier utilized in this AL study. Theoretical descriptions of this implementation accepting weights for the training instances are presented in Nguyen et al. [2010] for classification purposes as well as in Chang et al. [2004] for regression tasks. We will now detail the main points differentiating it from the standard version of the SVM outlined in Section 3.3.

Weighted primal and dual SVM

During the optimization of the weighted variant of the SVM, one assigns sample weights $\boldsymbol{\omega} = \{\omega_i\}_{i=1}^{n}$, $\omega_i \in \mathbb{R}^+$ to all the n training samples belonging to the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$. Then, the training of the weighted

SVM implies solving the primal problem of Eq. (3.6) (see page 35) modified as

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{n} \omega_i \xi_i \right\} , \tag{6.1}$$

but subject to the same constraints (3.7) and (3.8). The associated dual problem of Eq. (3.9) remains the same except for the condition (3.11) which becomes

$$0 \leq \alpha_i \leq \omega_i C \qquad \forall i , \tag{6.2}$$

where the α_i are the Lagrange multipliers related to each training point in the final (linear) SVM decision function (3.12), which also remains intact.

Changed upper-bound for the α_i

One can notice the upper-bound for such coefficients defining the actual influence of the SVs (training points with $\alpha_i > 0$) being dependent on the sample weight ω_i. This induces an increased flexibility of the method, with samples allowed to receive α_i coefficients larger than the employed C value when $\omega_i > 1$ (see discussion of the last paragraph of Section 3.3.1). Consequently, particularly relevant samples could have an additional impact on the classification system if compared to the usual SVM implementation.

6.2.2 *TrAdaBoost and Active Learning*

Notation used in this Chapter

Throughout this Chapter, sample pairs of both domains will actually be denoted with $(\boldsymbol{x}_i, y_i)$. The adaptive AL procedure detailed here will use a joint training set $T = T_S \cup T_T$ composed of a source training subset T_S and a target training subset T_T. Source samples pairs $(\boldsymbol{x}_S, y_S) \in T_S$ will be indexed as $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_S}$, whereas target samples $(\boldsymbol{x}_T, y_T) \in T_T$ will be indexed as $\{(\boldsymbol{x}_i, y_i)\}_{i=n_S+1}^{n_S+n_T}$. Additionally, the total number of labeled samples at each stage of the AL procedure will be designated by $n = n_S + n_T$.

Approach with two nested loops

To achieve DA through AL, two nested loops are run in order to select the most useful samples in the target image (outer AL loop) while iteratively adapting the resulting classifier to the new domain (inner TrAdaBoost loop). The scheme of Fig. 6.1 outlines the general procedure while Algorithm 1 provides details about its main steps. In the following, the two phases of the algorithm are described and their objectives are highlighted.

Initialization

Initially, the available labeled training set T is composed of the n_S source samples only, i. e. $T = T_S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n_S}$, because the target training set is initialized as $T_T = \{\}$, i. e. $n_T = 0$. Afterwards, T_T is progressively extended by appending each time q target training instances selected via AL. To this end, we provide the active learner with a set of unlabeled target domain candidates $U = \{\boldsymbol{x}_j\}_{j=1}^{n_U}$ among which to choose the interesting samples to be labeled by the user. Moreover, we initialize training sample weights as $\omega_i = 1 \,\forall\, i$. We employ this initialization instead of that with uniform weights $\omega_i = \frac{1}{n}$ [Dai et al., 2007], to let the second term of (6.1) become $C \sum_{i=1}^{n} \xi_i$, as in the usual SVM formulation. The role of the two nested loops is as follows.

AL loop

1. The outer loop of the adaptation procedure is an AL routine where, at each iteration, the q most interesting candidates $\mathbf{x}_j \in U$ are identified using the BT strategy and, after the assignment of the corresponding true label y_j, added to T_T. This heuristic selects the best points $\hat{\mathbf{x}}^{BT}$ according to the following ranking criterion [Luo et al., 2005]:

$$\begin{aligned}\hat{\mathbf{x}}^{BT} &= \arg\min_{\mathbf{x}_j \in U} \Big(\max_{cl \in \mathcal{C}} p(y_j^* = cl | \mathbf{x}_j) - \\ &\max_{cl \in \mathcal{C} \setminus cl^+} p(y_j^* = cl | \mathbf{x}_j) \Big), \end{aligned} \tag{6.3}$$

where $cl^+ = \arg\max_{cl \in \mathcal{C}} \big(p(y_j^* = cl | \mathbf{x}_j) \big)$ is the class with the highest probability for pixel $\mathbf{x}_j$ and $\mathcal{C} = \{1, 2, \dots, c\}$ is the set of c classes. These posterior probabilities are the output of the SVM classifier weighting the samples by means of vector $\boldsymbol{\omega} = \{\omega_i\}_{i=1}^n$ and are estimated with the Platt's method [Platt, 1999]. After the inclusion of the best candidate points to T_T, the complete training set T is updated as $T = T_S \cup T_T$.

TrAdaBoost loop

2. At each AL iteration, the inner TrAdaBoost loop is run to reweight the training instances in T. After having added the new labeled training samples, at the boosting iteration $t = 0$ we initialize a new weighting vector $\mathbf{v}^t$ by setting equal weights $v_i^0 = 1$, $\forall\, i$. Then, for every round of the inner loop, we consider the labels y_i^* predicted by the current SVM model for the training samples. In the multi-class case (extension of the binary problem approached in Dai et al. [2007]), the weighted training error on the target set T_T is then computed as:

$$\epsilon_t = \sum_{i=n_S+1}^{n_S+n_T} \frac{v_i^t \cdot e_i}{\sum_{i=n_S+1}^{n_S+n_T} v_i^t} \tag{6.4}$$

where e_i takes a value of 1 if the classifier commits an error ($y_i^* \neq y_i$) when labeling $\mathbf{x}_i$ and 0 otherwise ($y_i^* = y_i$). Afterwards, the weights v_i^t are updated for the subsequent boosting iteration in two distinct ways according to the domain of origin of $\mathbf{x}_i$. In fact, we apply

$$v_i^{t+1} = \begin{cases} v_i^t \beta^{e_i} & \text{if } \mathbf{x}_i \in T_S \\ v_i^t \beta_t^{-e_i} & \text{if } \mathbf{x}_i \in T_T\,, \end{cases} \tag{6.5}$$

where

$$\beta = 1/(1 + \sqrt{2 \ln n_S / tmax})\,, \tag{6.6}$$

$$\beta_t = \epsilon_t / (1 - \epsilon_t)\,. \tag{6.7}$$

The process is run for $tmax$ iterations and the final weights $\mathbf{v}^{tmax}$ are used to retrain the instance weighting SVM ($\boldsymbol{\omega} = \mathbf{v}^{tmax}$), yielding the predictions in the target domain (test set and unlabeled candidates set). The associated estimated class probabilities are subsequently used by BT in the AL loop to perform the active selection on the pool of candidates U.

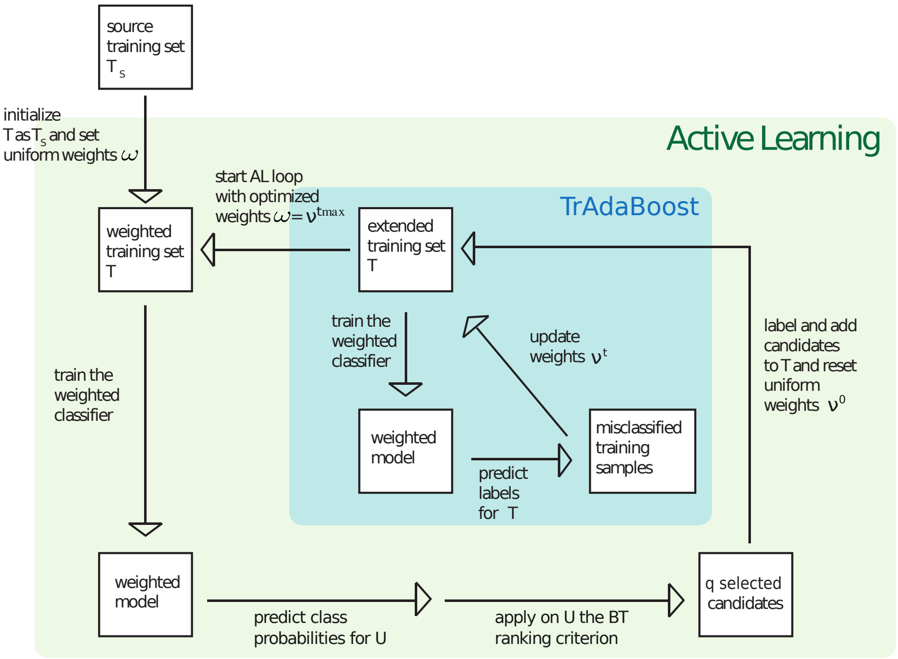

source training set T_s
initialize T as T_s and set uniform weights ω
Active Learning
start AL loop with optimized weights $\omega = \nu^{tmax}$
weighted training set T
extended training set T
TrAdaBoost
train the weighted classifier
update weights ν^t
label and add candidates to T and reset uniform weights ν^0
train the weighted classifier
weighted model
misclassified training samples
predict labels for T
weighted model
predict class probabilities for U
apply on U the BT ranking criterion
q selected candidates

Figure 6.1: Scheme of the adaptive AL procedure: the outer AL loop is highlighted in green, while the inner TrAdaBoost loop is highlighted in blue.

Algorithm 1 Adaptive AL with TrAdaBoost

```
Inputs: initial labeled source training set $T_S = \{(x_i, y_i)\}_{i=1}^{n_S}$, set of unlabeled
target domain candidates $U = \{x_j\}_{j=1}^{n_U}$, number of candidates to add at each
iteration $q$, number of TrAdaBoost iterations $tmax$
initialize $T_T = \{\}$, i. e. $n_T = 0$
initialize $T = T_S$
initialize $\boldsymbol{\omega}$ with equal unit weights $\omega_i = 1 \, \forall \, i$
for each AL iteration do
    train the SVM using $T$ (weighted by $\boldsymbol{\omega}$) as training set
    compute the SVM test accuracy in the target domain
    predict the $c$ class probabilities $p(y_j^* = cl | \mathbf{x}_j) \, \forall \, \mathbf{x}_j \in U$
    compute ranking criterion according to Eq. (6.3)
    remove the best $q$ candidates from $U$ and add them to $T_T$
    set $T = T_S \cup T_T$
    set $\boldsymbol{\nu}^t = \boldsymbol{\nu}^0$ to unit weights $\nu_i^t = \nu_i^0 = 1 \, \forall \, i$
    for each TrAdaBoost iteration $t = 1, \ldots, tmax$ do
        train the SVM (weighted by $\boldsymbol{\nu}^t$) using the extended $T$
        repredict the class labels $y_i^* \, \forall \, \mathbf{x}_i \in T$
        calculate the weighted error $\epsilon_t$ on $T_T$ according to Eq. (6.4)
        update weights to obtain $\boldsymbol{\nu}^{t+1}$ following Eq. (6.5)
    end for
    set $\boldsymbol{\omega} = \boldsymbol{\nu}^{tmax}$
end for
Outputs: final training set $T$, test classification accuracy along the AL iterations
```

Update of the sample weights

Taking a closer look at the TrAdaBoost loop, in Eq. (6.5) one will notice that if the sample is correctly classified, the weight remains unchanged, whereas if the sample is misclassified, two options are possible. If the sample comes from the source domain, its weight is decreased by a constant factor (6.6). Conversely, if the instance originates from the domain of interest, the target domain, its weight is increased by a factor inversely proportional to the target training error (6.7). This updating strategy aims at reducing the impact of misleading source examples, supposed to be the most dissimilar to the target instances the model should focus on. Conversely, the increase of the influence of misclassified target samples translates the need to concentrate on the regions of the target domain in which the class discrimination is harder. In light of these considerations, the boosting loop could be prone to overfit potential outliers. However, let us remark that, when the weighted target training error ϵ_t is excessively large (> 0.5), the reweighting factor β_t exceeds the value of 1, allowing therefore a decrease of the weights for the misclassified target samples in (6.5).

Benefits of the procedure

This transfer learning approach enables the SVM model to gradually adjust itself to the new domain. The different weighting of the examples leads to a boosted decision function more and more suited to model the input-output relationships in the target domain. Hence, the benefits of this procedure are twofold. On the one hand, the quality of the classification on test data (belonging to the target domain) is improved. On the other hand, since we are acquiring samples representing the target distribution, the class membership probabilities for the unlabeled samples in U are more accurately computed. This induces an AL selection criterion better suited to identify candidates lying in uncertain regions of the extended input space in the following iterations.

6.3 DATA AND EXPERIMENTAL SETUP

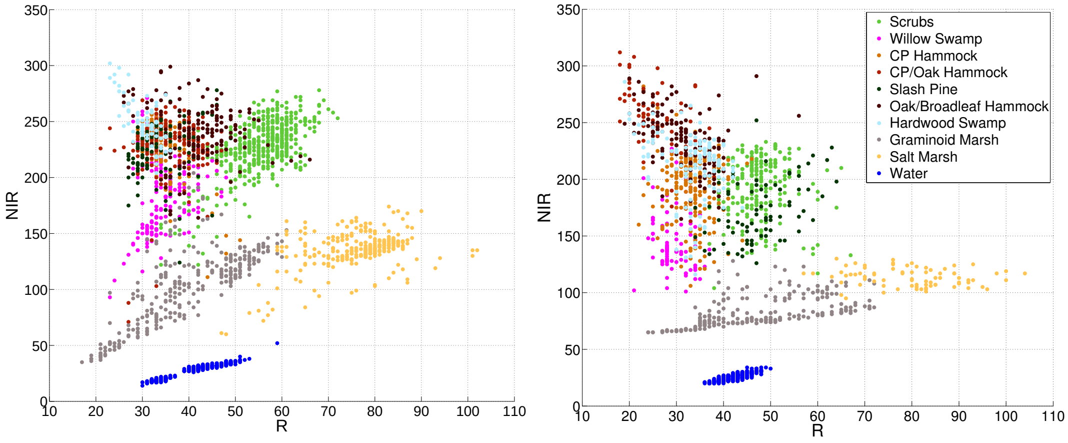
Two case studies

In the following Sections, we describe the data we considered for the experiments as well as the related setup. The proposed methodology has been tested on two datasets. The first one represents an urban case study bearing a moderate shift between the source and target images. On the contrary, in the second dataset the target domain is represented by a region showing remarkable differences in the spectral signatures of the vegetative cover with respect to the source region.

6.3.1 *VHR QuickBird images of Zurich*

Images and pre-processing

The first dataset consists of the two VHR QuickBird images of the city of Zurich (Switzerland) presented in Appendix B.1 (see page 152). The histograms of the two images have first been matched via HM (see page 25) taking the source image as the reference and, subsequently, textural (3 × 3 data range, mean, homogeneity and entropy) [Haralick et al., 1973] and morphological (5 × 5 opening and closing, 7 × 7 and 9 × 9 opening and



(a)

(b)

Figure 6.2: Scatterplots for the two KSC images in the red (AVIRIS band #29: ≈ 667 nm) vs. NIR (AVIRIS band #49: ≈ 831 nm) space. (a) Source training set. (b) Target test set. In the legend: CP = cabbage palm.

closing by reconstruction) [Pesaresi and Benediktsson, 2001] features have been extracted from the panchromatic band to enrich the ground-cover description with spatial information. For the following experiments, the total number of considered features is thus 15 (4 VNIR bands, 1 PAN, 4 textural, 6 morphological). Prior to the analyses, the variables have been normalized to have zero mean and unit variance, based on the source image descriptive statistics.

Ground truth and datasets

The ground truth, whose details are listed in Tab. B.1 of the Appendix, is made up of thematic classes usually encountered in urban environments. For the target image we derived two separate datasets: the set of unlabeled samples and the test set. On the one hand, the set of candidates U to be provided to the AL procedure (assuming their true label unknown) includes 22,723 pixels. On the other hand, the generalization ability of the different techniques in the target domain has been assessed on 26,797 test samples issued from spatially separate regions of the target image. For the source image, only a labeled set was needed and was composed of 15,934 pixels.

6.3.2 *Hyperspectral AVIRIS image of the KSC*

Images and pre-processing

The second case study deals with land-cover classification in a subtropical region. The source and target images have been defined as sub-regions of the same hyperspectral image acquired by AVIRIS over the *Kennedy Space Center* (KSC), Florida (USA). For the details related to this dataset we refer the reader to Appendix B.2 (see page 154). As for the QuickBird dataset, after a matching of the histograms, the bands have been normalized (zero mean and unit variance) using source image parameters.

Ground truth and datasets

The list of the classes to discriminate during the experiments, mostly subtropical vegetation land-cover, is provided in Tab. B.2. As depicted by the scatterplots of Fig. 6.2, the classes have a rather large overlap along with important radiometric variations across domains. The resulting dataset shift observed from one image to the other is marked (divergence in both class-conditional and marginal probability distributions). A labeled set consisting of 2,522 pixels was issued from the source image. For the target image, we partitioned the available labeled dataset into an unlabeled set of candidates U and a test set both including 1,927 pixels. This independent target test set is then used for the comparison of the performances of the AL strategies.

6.3.3 *Experimental setup*

Size of training sets, classifier & parameters

The experiments were conducted with 10 different and independent realizations of the initial source training sets T_S (drawn from the labeled sets mentioned above). For the Zurich images $n_S = 1000$ randomly selected pixels were retained, while the size of the training set was fixed to $n_S = 500$ for the KSC dataset. An instance weighting SVM with a linear kernel has been used as supervised learner and a 5-fold cross-validation has been performed to find the optimal initial C parameter (extensive search in the space $\{10^{-1}, \ldots, 10^5\}$). For both datasets and for all the AL methods, $q = 10$ target samples per iteration were added to augment the initial source training set while the AL process was run for 35 iterations. At each iteration, the performance of the SVM models has been assessed on the test set extracted from the corresponding target image.

Compared methods

We compared the proposed adaptive AL strategy (`AdaptiveAL_BT`) with the standard BT without instance reweighting (`AL_BT`) and with a procedure randomly selecting the pixels to label in the target image while adapting their weights following the TrAdaBoost scheme (`AdaptiveRandomS`). Also, in order to provide the usual AL baseline, the random selection of the samples to label (`RandomS`) has been considered. Finally, to set reference performances for the considered target images, linear SVM classifiers exclusively trained on source (same 10 independent samplings of n_S pixels from the complete training set) and target (also n_S pixels sampled from the set of candidates U retaining their labels) datasets have been tested (`Source` and `Target` methods, respectively). Regarding the proposed `AdaptiveAL_BT` method and `AdaptiveRandomS`, at each AL iteration, the weights of the samples in the training set were updated after 5 iterations of TrAdaBoost (stabilized v_i values). In this sub-routine, the prediction on the training set was implemented through a 20-fold cross-validation to avoid overfitting.

Software

The algorithms were implemented in MATLAB® using LIBSVM as library both for the standard SVM and instance weighting SVM (version available at `http://www.csie.ntu.edu.tw`) [Chang and Lin, 2001]. The computation of class probabilities to be used by BT is described in the same paper.

6.4 RESULTS

6.4.1 *Learning curves*

Figure 6.3 summarizes the results for this task of DA through AL. The performance of the different AL techniques along the iterations (increasing training set size) has been assessed in terms of overall classification accuracy (see Appendix A). The depicted learning curves represent the average OA over the 10 experiments.

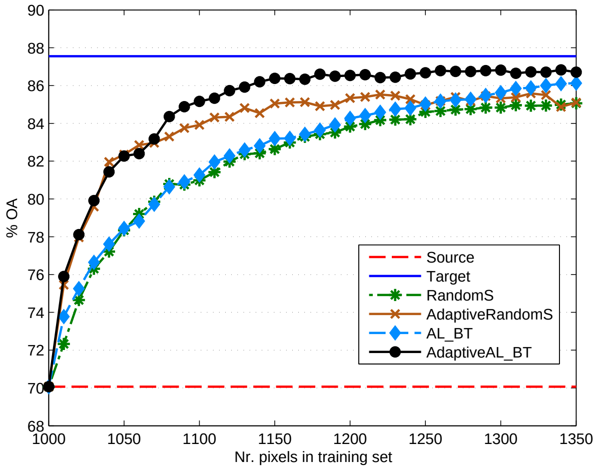
6.4.1.1 *VHR QuickBird images of Zurich*

Analyzing Fig. 6.3(a) referring to the Zurich dataset, one can first notice the bad performance achieved by applying on the target data the source model (`Source`) without any adjustments (OA = 70.07%). The method consisting in randomly sampling the pool of unlabeled pixels (`RandomS`), considered as a baseline for AL, and the standard AL heuristic of BT both reveal a slow convergence. Nevertheless, the `AL_BT` method yields SVM models that are slightly more accurate than those built by sampling at random, but this happens only from the 10th iteration onwards (approximately +0.5% OA).

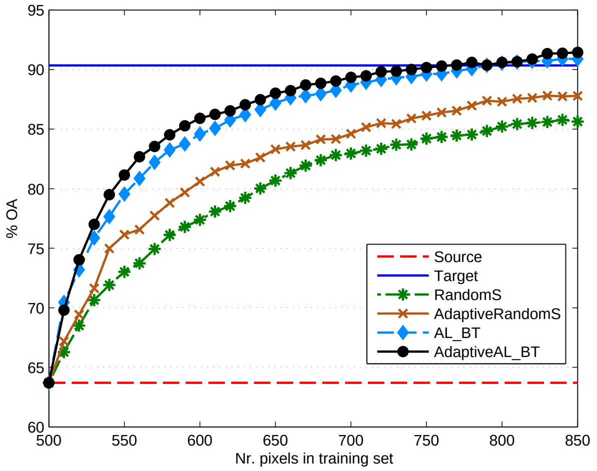
The proposed combined methodology integrating the TrAdaBoost routine in the AL process (`AdaptiveAL_BT`) clearly outperforms these two sampling schemes by sharply increasing the classification accuracy since the very beginning of the AL iterations. In fact, already after 14 iterations (140 target pixels added) the associated curve achieves an OA of 86.2% (+3.4% OA with respect to `AL_BT`). Such a precision is never reached by the two baseline approaches during the considered first 35 AL cycles. Nevertheless, we remark how the other procedure including the reweighting scheme, `AdaptiveRandomS`, is yielding a performance comparable to that of its active counterpart in the first 7 cycles of the AL routine. Subsequently, after the addition of 80 samples, the actively guided selection of the pixels to label provides an average improvement in OA of 1.5%. It is interesting to note that none of the strategies is able to reach the `Target` performance at OA = 87.55%.

6.4.1.2 *Hyperspectral AVIRIS image of the KSC*

Figure 6.3(b) reveals a similar pattern in the KSC dataset, except for the improved performance of the `AL_BT` strategy and a worsened performance of the `AdaptiveRandomS` method. The random sampling of the pixels in the target image (`RandomS`) results in poor updates of the initial training set. In fact, even after the inclusion of 350 samples, the model still lies 5% OA below the performance of a SVM trained with pixels from the target image only (`Target` reference classification with average OA = 90.35%). On the other hand, both the `AdaptiveAL_BT` and the `AL_BT` show promising learning curves, eventually reaching and even exceeding the upper reference accuracy of the same-domain SVM model. In particular, one can remark the



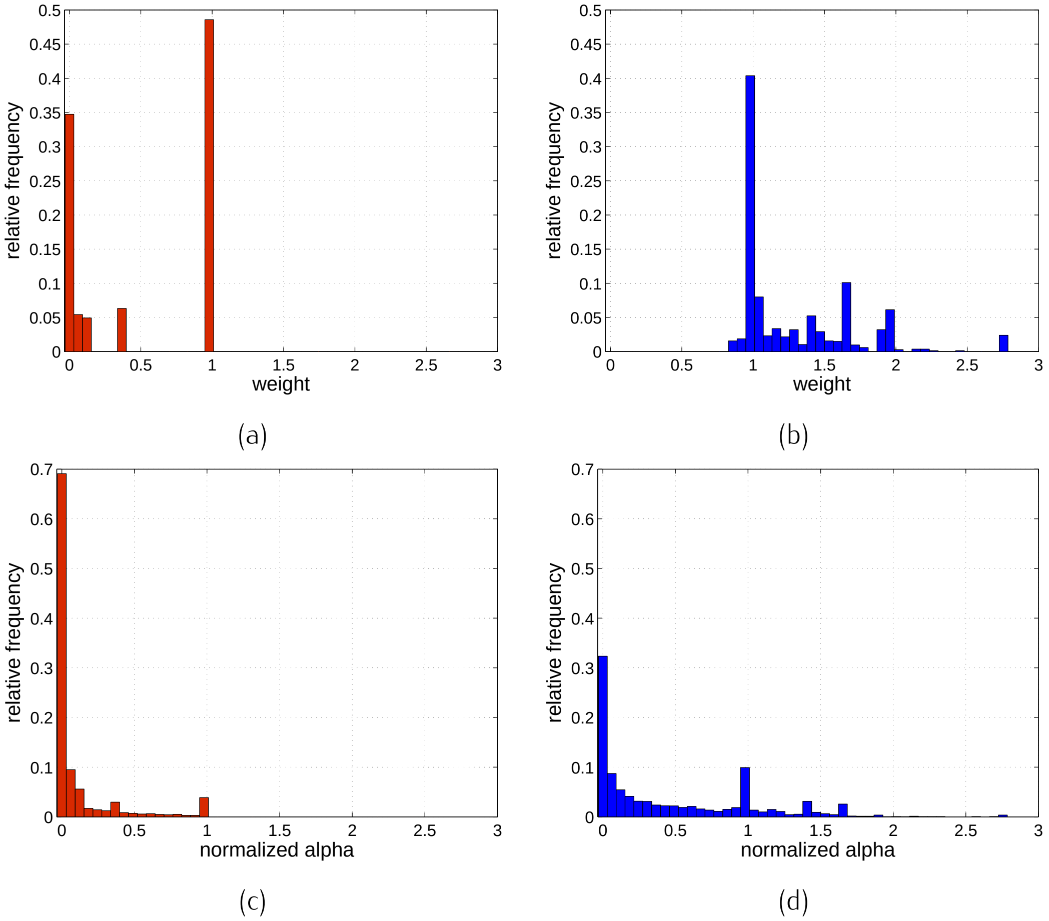
(a)



(b)

Figure 6.3: Average learning curves (% OA) over 10 runs on the target image of the (a) Zurich dataset and (b) KSC dataset. `Source` (dashed red line) = model built using pixels of the source domain only, `Target` (solid blue line) = model built using pixels of the target domain only, `RandomS` (dashed green line with asterisks) = random sampling method, `AdaptiveRandomS` (solid brown line with crosses) = random sampling method combined with TrAdaBoost, `AL_BT` (dashed light blue line with diamonds) = AL via breaking ties, `AdaptiveAL_BT` (solid black line with circles) = proposed adaptive AL method.

adaptive AL procedure evolving $\approx$ 1% OA higher than its non-adaptive counterpart. The effect of the intelligent selection of the most informative pixels, as provided by the BT strategy, when combined with the TrAdaBoost algorithm is more evident on the KSC dataset. In fact, the curve associated with the integration of TrAdaBoost with the random, passive sampling in the new image (`AdaptiveRandomS`) remains between 4% and 6% OA lower than the active one from the beginning of the AL process.



(a)

(b)

(c)

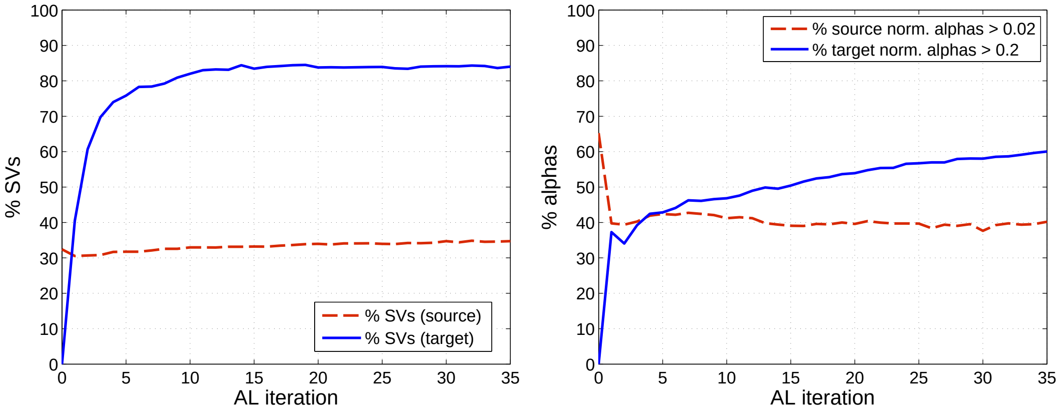
(d)

Figure 6.4: KSC dataset: histograms showing the distribution (frequencies over the 10 experiments) of the final TrAdaBoost weights ω_i and SVM coefficients α_i (normalized by C to account for its different values in the experiments) for the SVs of the two domains at the AL iteration #15 ($n = n_S + n_T = 500 + 150 = 650$ pixels in the training set). (a) Source SVs: weights ω_i. (b) Target SVs: weights ω_i. (c) Source SVs: α_i / C. (d) Target SVs: α_i / C.

6.4.2 *Analysis of sample weights*

With the purpose to shed light on the actual effect of the TrAdaBoost model on the SVM-based AL procedure, it is worth analyzing the evolution of the weights ω_i and the coefficients α_i along the AL iterations.

Figure 6.4 illustrates the distribution of the respective weights ω_i and coefficients α_i for the SVs of each domain at the 15th AL iteration. Indeed, these are crucial training samples, the only ones contributing to the final SVM decision function. With this example concerning the KSC dataset, we focus on the state of the `AdaptiveAL_BT` method. From Figs. 6.4(a) and 6.4(b), one can observe how the weights of roughly 60% of the training SVs belonging to the source image are set to very low values ($\omega_i < 0.15$), whereas more than half of those of target domain SVs take values larger than 1. This translates, for the source SVs (Fig. 6.4(c)), to a significant



(a) (b)

Figure 6.5: KSC dataset: evolution along the AL iterations of the ratio of SVs and magnitude of the alpha coefficients for the two domains (percentages over the 10 experiments). (a) Percentage of SVs among source (dashed red line) and target (solid blue line) training data. (b) Percentage of source (dashed red line) and target (solid blue line) normalized alphas (α_i / C) larger than 0.02 or 0.2, respectively (percentages computed over the total number of alphas obtained in each domain).

amount of alpha coefficients found to be close to 0 and, for the target SVs (Fig. 6.4(d)), to a non-negligible number of alpha values that are actually larger than the corresponding SVM hyper-parameter C, i. e. $\alpha_i / C > 1$.

The highlighted tendency is noticeable since the early stages of the AL cycle, with more importance given to useful instances in the target domain and, conversely, with less weight assigned to misleading source instances. To better perceive the cited evolution as the AL and TrAdaBoost loops proceed to the adaptation of the SVM, we resort to Fig. 6.5.

Figure 6.5(a) depicts the evolution of the share of training points that eventually become SVs in the two domains. The number of such key samples remains stable over the entire AL procedure for the source training set T_S. On the contrary, for the target training set T_T we notice a growth of the considered ratio of SVs which is especially steep at first (until iteration 4), and then gradually slows down as the new image is sampled.

In Fig. 6.5(b), it is insightful to notice how, among the alpha coefficients (represented by their normalized counterparts α_i / C) associated with the SVs, there is a consistent polarizing trend as the AL algorithm runs. In fact, always more and more of these α_i take either high values if corresponding to target samples, or low values if representing source samples. This evolution of the alphas is more marked for the target image, almost doubling the proportion of normalized $\alpha_i > 0.2$ found in the first iterations by the time the AL loop reaches its end. It is worth pointing out the sheer drop (from 65.2% to 39.7%) in the proportion of source normalized alphas larger than 0.02 when the first $q = 10$ target samples are added to the joint training set T. The model reweights samples according to the domain of origin and this is reflected in the SVM coefficients.

6.4.3 *Discussion*

As pointed out in Section 6.4.1, an appropriately designed weighting scheme for the training instances, as the one provided by the presented method, ensures an improved transfer of the knowledge between the source and the target image. A direct consequence of this fact, due to the improved posterior class probability estimates, is the more accurate selection of samples to be labeled along the subsequent AL iterations. We obtain an improved model, able to outperform in test the one built by selecting the training instances with the simple BT heuristic, if the latter is naively applied without any adaptation to the domain of interest. Moreover, improvements over the simple application of the TrAdaBoost algorithm in combination with a random sampling of the new image were observed. This highlights the impact of the active selection, via BT in this case, of the most helpful pixels of the target image.

In more detail, we can comment on the influence of the two qualities an adaptive AL system should possess: the ability to adapt to the domain of interest and the ability to actively select the new samples. The experiments we conducted reveal an opposite trend in the two considered datasets. On the one hand, on the Zurich images, we notice a higher importance given to the adaptation to the new domain (superior performance of the **`AdaptiveRandomS`** over the **`AL_BT`** method). That could be linked to the need of downweighting source pixels found in areas related to the shift, but in a rather stable environment, in terms of marginal distributions. At the same time, the misclassified target pixels, lying in a region where the class boundaries have changed, require more attention (increasing weights) to adjust the model. On the other hand, when dealing with the KSC dataset, the effect of the active sampling alone proves to be more decisive than the simple adaptation of the weights (superior performance of the **`AL_BT`** method). This behavior can be linked to the larger and nonlinear shift observed in this second hyperspectral case, that forces the algorithm to completely redefine the decision boundaries with the new queries. These additional samples are extremely useful to precisely redefine the new distribution of the highly mixed and overlapping classes that characterize the study area.

Despite the contradictory behavior observed in the case studies (in the first DA is more beneficial than AL, while in the second the opposite holds), the proposed method returns the best results in both cases. First, this illustrates the complementarity of the AL and DA approaches, that are effective in different scenarios. Second, this also strengthens the interest of a joint approach capable of taking the best from both worlds: in the nested loops of the proposed strategy, AL and DA interact constantly and can thus provide the relevant samples, while adapting the model to the new domain. The consequence is the remarkable gain in classification accuracy during the first iterations, observed in both case studies when using adaptive active sampling strategies.

Additionally, it is worth noting that, for the KSC images, both the active strategies converge to a performance exceeding that of the model built exclusively on target data. These superior classification accuracies are obtained with training sets that required the labeling of 270-290 target pixels only and thus showing the interest of intelligently built compact models avoiding the labeling of redundant samples. Furthermore, this fact indicates that the source data is still relevant and brings into play universal information that is useful to solve the problem in the target domain. This accuracy improvement is even more significant in light of the large shift of the class spectral signatures existing between the two images, as testified by the ≈ 26.6% OA difference between the `Source` and `Target` models (see also scatterplots of Fig. 6.2).

Section 6.4.2 emphasizes instead the usefulness and the impact of the dedicated instance weights included in the SVM model, core of the proposed adaptive AL approach. As pointed out in Section 6.2.1, the standard kernel-based learning machine optimizing the alpha coefficients with an equal upper-bound ($\alpha_i \leq C, \forall\, i$) is turned into an adaptable learning machine ($\alpha_i \leq \omega_i C, \forall\, i$). This weighted version of the SVM, as a matter of fact, is able to accord distinct relevance values to the training examples following both their domain of origin and their contribution to the class discrimination task. We draw the attention on the fact that, since these alpha coefficients act as sample weights in the SVM final decision function (3.12), the predictions are notably affected by the TrAdaBoost reweighting scheme.

In this sense, the evolution curves of Fig. 6.5 testify the increasing influence on the classification system of the pixels collected in the target image. As batches of these new domain samples are included in the training set, they quickly display a higher likelihood to become SVs than the already present source samples. Furthermore, the magnitude of the associated alpha coefficients is also increasing, translating the augmented relevance of the pixels belonging to the target domain we are interested in. The adjustable instance weights boost the SVM performance and enable the model to assign tailored alpha coefficients to its SVs. The AL process efficiently adapts the classifier by attributing more and more importance to the target domain while discarding unprofitable source information. As a result, we obtain an improved discrimination of the land-cover classes in the image for which we need to produce a new thematic map.

6.5 CONCLUSIONS

Main achievements

In this Chapter, an approach to boost the performance of AL methods when applied in the context of DA has been presented and analyzed. We described a technique, TrAdaBoost, aimed at properly adapting training sample weights during the AL process. Such adjustments proved potential in refining the ranking criterion for the selection of the most informative target pixels to be manually labeled by the user. The individual contributions of the smart sampling and of the adaptive adjustment of sample weights have been assessed, concluding that the best performances are obtained when the two approaches are combined.

Influence of the reweighting

One of the objectives was also to uncover and better understand the behavior of the proposed reweighting scheme when integrated with a SVM classifier accepting instance weights in the training phase. The influence of these weights on the decision function, conveying the importance and pertinence to the domain of interest of each pixel, has been highlighted through the analysis of the evolution of the SVs all along the sampling procedure.

Benefits in concrete applications

The present Chapter demonstrated that in a classification task involving a newly acquired image and when already collected ground truth data are available, the modeling effort for the target image can be efficiently reduced. In fact, by means of the proposed adaptive sampling strategy, the operator will be properly guided in the collection of the labels for the most useful pixels on the new image. Standard supervised classifiers are supplied with a minimal and effective training set for a suitable land-cover thematic mapping.

Future work

Further developments of adaptive AL approaches could concentrate on one of the open issues that have been seldom addressed in the literature so far: the change in the set of classes from one image to another. Indeed, in truly large-scale applications, the common DA assumption that the types of land-cover are the same across the entire study region is often violated. On the one hand, intelligent sampling strategies able to discover new thematic classes previously unseen in the source image are in high demand. On the other hand, attention should also be paid to appropriately handle the disappearance of a given land-cover when moving to the target image.

7 KERNEL-BASED FEATURE EXTRACTION FOR RELATIVE RADIOMETRIC NORMALIZATION

Outline: *In this Chapter, we study the problem of Feature Extraction for the relative normalization of multiple remotely sensed images in order to ease the knowledge transfer among them. We analyze a recently proposed Feature Extraction method specifically designed for Domain Adaptation, Transfer Component Analysis, and its semisupervised implementation named Semisupervised Transfer Component Analysis. In Section 7.1, we briefly recall the motivation for this work and in Section 7.2, we formalize the associated Feature Extraction framework. Next, in Section 7.3 we present the studied techniques. Section 7.4 describes the datasets used and the setup of the experiments whose results are presented and discussed in Section 7.5. Finally, in Section 7.6, we summarize the main findings of this study while providing further possible research directions.*

7.1 INTRODUCTION

Benefits of FE for DA

Previously, in the introductory Part i of this Thesis, we have seen how the issues affecting the radiometry of the images acquired under different conditions (see Section 2.5.2) can be alleviated by adopting image normalization strategies. Among these compensations methods briefly reviewed in Section 2.5.3, we find relative normalization techniques that aim at providing a dedicated image-to-image calibration. It is indeed in this context that the application of the suitable feature-representation-transfer methodologies can reveal potential (see Section 5.2.3). Particularly attractive are the alignment strategies based on the extraction of totally new features derived from the initial data with the purpose to bridge the gap existing between the images. In this context, if the only labeled ground truth data refers to the source image, FE can be considered as a means of resolving the image adaptation problem in an unsupervised way. Both images are mapped in a common latent space where the class boundaries are expected to be

This Chapter is based on the following accepted paper:

> G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia. Semisupervised Transfer Component Analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, Accepted.

.

more domain-invariant and, subsequently, a prediction based on the source training set can be performed.

Overview of the Chapter

In the present Chapter, we study the effectiveness of non-linear FE techniques when applied in a cross-domain setting where we have at our disposal labeled samples only in the source image. In particular, the novelty of the work consists in the investigation of the capabilities of a FE method especially developed for DA. We consider the recently proposed Transfer Component Analysis and, specifically, we focus on its extension named Semisupervised Transfer Component Analysis [Pan et al., 2011]. Both approaches explicitly minimize a term measuring the distance between the domains: the MMD presented in Section 4.3.2. Moreover, SSTCA also includes in the objective a manifold regularization term enforcing the smoothness of the projections and a label dependence term. The former enforces the preservation of the local geometry (data manifold) while the latter maximizes the alignment of the projections with the available source domain labels.

Contributions

We analyze these methods in a range of settings specifically designed to enforce the similarity of the domains. In particular, we study in detail the behavior of SSTCA with respect to its key parameters and related objectives of the projection. A thorough comparison to a number of general purpose feature extractors which may also be used in a DA setting is provided. First, the quality of the alignment is assessed in classification tasks involving spatially disjoint pairs of images acquired by multi- and hyperspectral sensors. Then, we perform a visual assessment of both the *invariance property* and the class *discrimination property* of the extracted features. Furthermore, several other key issues related to the cross-image knowledge transfer by FE are studied. First, we analyze the combined use of the considered FE strategies with the widely utilized HM procedure. Second, we evaluate the influence of the origin of the unlabeled samples used for FE (source image only or both source and target images). Third, we assess the importance of spatial features by comparing the DA results obtained using spectral-spatial information to those using spectral data only.

7.2 DOMAIN ADAPTATION VIA FEATURE EXTRACTION

The principle

As introduced in Section 4.4.2, the purpose of feature-representation-transfer strategies is to find a common representation of the source and target datasets that minimizes the differences between these two domains while maintaining their main data properties (data statistics, local relationships, label dependence, etc.). Once the samples are mapped to the same subspace defined by the extracted features, a classifier is trained in the source domain using the available labeled examples, and then inference is performed directly in the aligned target domain.

Problem definition

Let us consider the set of n_S labeled source training data $D_S = \{X_S, Y_S\} = \{(\mathbf{x}_{S_i}, y_{S_i})\}_{i=1}^{n_S}$ and the set of the n_T unlabeled target data $X_T = \{\mathbf{x}_{T_j}\}_{j=1}^{n_T}$. The goal of the unsupervised DA approach considered in this Chapter is to predict target labels $y_T \in \mathcal{C} = \{1, 2, \ldots, c\}$ (set of c classes in common

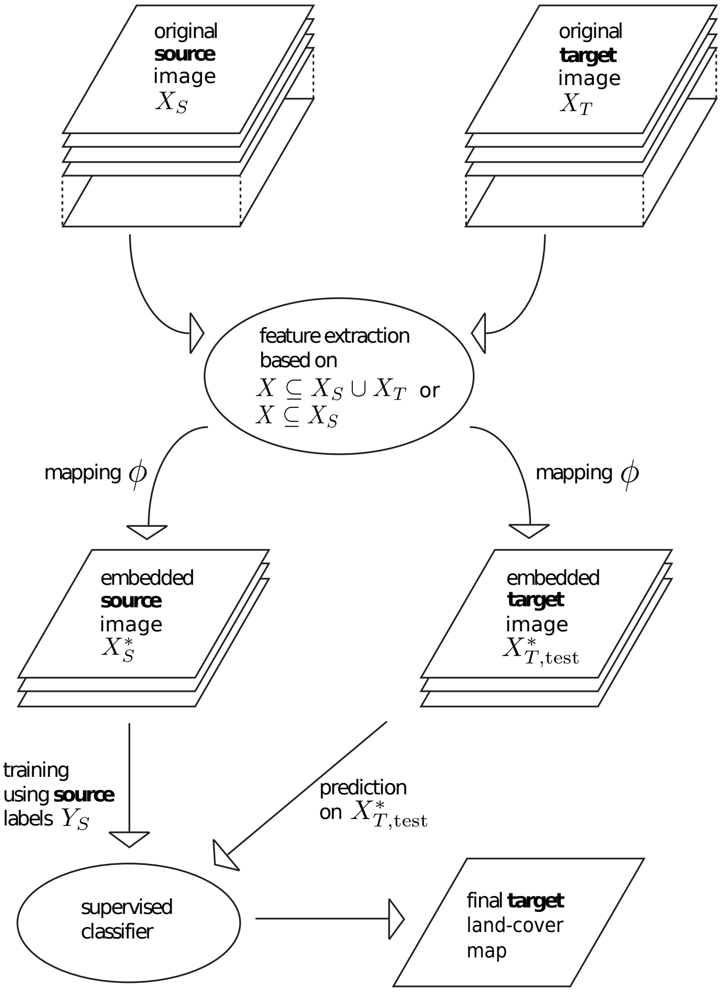


Figure 7.1: Flowchart of the considered FE approach to DA in image classification.

with D_S) based exclusively on the use of labeled data from D_S in the training phase. To this end, a mapping ϕ of the samples of both domains to a common space is needed: $X_S \to \phi(X_S) = X_S^*$, $X_T \to \phi(X_T) = X_T^*$. After this projection the two domains should be aligned, i. e. their probability distributions should be as similar as possible: $P(X_S^*) \approx P(X_T^*)$. In practice, we need a matrix $\boldsymbol{W}$ to perform the joint mapping ϕ of the data. The projection matrix can be found using a set of samples X collected either from

- the two domains ($X \subseteq X_S \cup X_T$), or
- one domain only, ($X \subseteq X_S$ or X_T).

Various FE methods (see Section 3.5) can be employed to estimate this matrix embedding the data in a m-dimensional sub-space with $m \ll d$. Ultimately, the purpose is to train a classifier on the projected training set $\{(\boldsymbol{x}_{S_i}^*, y_{S_i})\}_{i=1}^{n_S}$ and then to apply it to predict class labels y_T for the projected target test set $X_{T,\text{test}}^*$ (the entire image). Figure 7.1 illustrates the concept of the considered FE-based DA.

7.3 TRANSFER COMPONENT ANALYSIS

Hereafter, we describe the non-linear kernel-based FE technique enabling the domain-approaching projection, the unsupervised TCA. Next, we outline its semisupervised extension, SSTCA, a development taking into account the need both for a preservation of the local manifold structure and for the alignment with the class labels.

7.3.1 *Unsupervised Transfer Component Analysis*

Objectives of TCA

The TCA technique we investigate in this Chapter has specifically been developed for DA [Pan et al., 2011], and is based on the MMD metric presented in Section 4.3.2. TCA aims at finding a common embedding of the data from the two domains by:

1) minimizing the distance between the probability distributions of $\phi(X_S)$ and $\phi(X_T)$ (MMD minimization),

2) preserving the main statistical properties of the original data X_S and X_T (maximization of data variance in the first extracted orthogonal components).

The mapping function ϕ is empirically estimated by a transformation matrix $\boldsymbol{W} \in \mathbb{R}^{(n_S+n_T)\times m}$ with $m \ll (n_S + n_T)$ which explicitly incorporates both objectives.

MMD minimization

Using the transformation matrix $\boldsymbol{W}$, and the original kernel matrix $\boldsymbol{K}$ of (4.7) (see page 51) built on the stacked source and target sets, it is possible to compute the kernel matrix between mapped samples as $\boldsymbol{K}^* = \boldsymbol{K}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{K}$. Note that the projection of the samples in the transformed domains is obtained as $\boldsymbol{X}^* = \boldsymbol{K}\boldsymbol{W}$. Thus, this matrix $\boldsymbol{K}^*$ is evaluated by dot products of the mapped samples as $\boldsymbol{X}^*\boldsymbol{X}^{*\top}$, where $\boldsymbol{X}^*$ is the non-linearly transformed data matrix of size $(n_S + n_T) \times m$ having the projected samples $\phi(\boldsymbol{x}) = \boldsymbol{x}^*$ as its rows. Consequently, rewriting the MMD formula of Eq. (4.6), the distance between the mapped samples can be obtained as

$$\begin{aligned}\mathrm{MMD}(X_S^*, X_T^*) &= \mathrm{Tr}((\boldsymbol{K}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{K})\boldsymbol{L}) \\ &= \mathrm{Tr}(\boldsymbol{W}^\top\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}\boldsymbol{W})\ . \end{aligned} \tag{7.1}$$

The goal stated in 1) is thus achieved by minimizing (7.1) with respect to $\boldsymbol{W}$, guaranteeing therefore that the distributions of the projected domains minimize the MMD.

Maximization of data variance

On the other hand, objective 2) requires that ϕ does not inappropriately deform the input space, which would complicate the task for the classification routine. Hence, matrix $\boldsymbol{W}$ should be found such that the projection into the newly created subspace is able to preserve (and ideally compress in few components, as in PCA or KPCA) the initial data variance. The covariance matrix Σ^* of the projected samples $\boldsymbol{x}^*$ is given by

$$\begin{aligned}\Sigma^* &= \frac{1}{n_S + n_T} \sum_{i,j=1}^{n_S+n_T} (\boldsymbol{x}_i^* - \bar{\boldsymbol{x}}^*)(\boldsymbol{x}_j^* - \bar{\boldsymbol{x}}^*)^\top \\ &= \boldsymbol{W}^\top \boldsymbol{KHKW} \,, \end{aligned} \tag{7.2}$$

where $\bar{\boldsymbol{x}}^*$ is the average of the mapped samples and $\boldsymbol{H} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^\top/(n_S + n_T)$ is the $(n_S + n_T) \times (n_S + n_T)$ centering matrix. Thus, the following orthogonality constraint will be integrated in the optimization problem (diagonal covariance): $\Sigma^* = \boldsymbol{I}$, where $\boldsymbol{I}$ is the $m \times m$ identity matrix.

TCA kernel learning problem

The final kernel learning problem for the unsupervised TCA is then set up as

$$\begin{aligned} \arg\min_{\boldsymbol{W}} \quad & \left\{ \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{KLKW}) + \mu \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{W}) \right\} \\ \text{s.t.} \quad & \Sigma^* = \boldsymbol{W}^\top \boldsymbol{KHKW} = \boldsymbol{I} \,, \end{aligned} \tag{7.3}$$

where μ is a tradeoff parameter tuning the influence of the regularization term $\mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{W})$ controlling the complexity of $\boldsymbol{W}$. The present optimization problem can be reformulated as a trace maximization problem whose final solution may be obtained by introducing Lagrange multipliers and setting the partial derivatives with respect to $\boldsymbol{W}$ equal to zero, thus solving the generalized eigenvalue problem [Pan et al., 2011]

$$\boldsymbol{KHKv} = \rho(\boldsymbol{KLK} + \mu \boldsymbol{I})\boldsymbol{v} \,. \tag{7.4}$$

This entails the eigendecomposition of $(\boldsymbol{KLK} + \mu\boldsymbol{I})^{-1}\boldsymbol{KHK}$. The mapping matrix $\boldsymbol{W}$ is obtained by stacking the m eigenvectors $\boldsymbol{v}_i$ associated with the m largest eigenvalues ρ_i of (7.4) as $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m]$.

Projection of new samples

Finally, we compute the m *transfer components* for new test samples $\boldsymbol{X}_{\text{test}}$ as $\boldsymbol{X}^*_{\text{test}} = \boldsymbol{K}_{\text{test}}\boldsymbol{W}$, where $\boldsymbol{K}_{\text{test}}$ represents the $n_{\text{test}} \times (n_S + n_T)$ kernel matrix between the n_{test} test points and the $(n_S + n_T)$ training samples.

7.3.2 *Semisupervised Transfer Component Analysis*

Additional objectives of SSTCA

Besides the regularized MMD minimization objective used by the unsupervised version of TCA, a desirable aligning projection should fulfill two additional requirements.

1) It should maximize the dependence between the extracted features and the class labels by exploiting the available source labeled samples (introducing a form of supervision in the definition of the appropriate projection).

2) It should preserve the local structure in both domains to further regularize the projection and to avoid an exaggerated deformation of the respective data manifolds with the joint transform.

Indeed, SSTCA pursues these two extra goals.

Label dependence

To achieve the first objective, we force the projections to be dependent on the available labeled data. A label indicator matrix is included in the objective function:

$$\boldsymbol{K}^*_{YY} = \gamma \boldsymbol{K}_{YY} + (1-\gamma)\boldsymbol{I} \, , \tag{7.5}$$

where $\boldsymbol{K}_{YY}$ is a kernel matrix computed on the labels Y that, while being of size $(n_S + n_T) \times (n_S + n_T)$, is only defined on the source domain: $K_{YY_{i,j}} = 1$ if $y_i = y_j$ with $\boldsymbol{x}_i, \boldsymbol{x}_j \in X_S$, whereas $K_{YY_{i,j}} = 0$ otherwise. The first term in (7.5) aims at maximizing the label dependence whereas the second serves to maximize the data variance in both domains. The two competing terms are balanced by the tradeoff parameter $\gamma \geq 0$. This alignment can be achieved by making use of a measure of dependence between sets of variables X' and Y' (the labels) called *Hilbert-Schmidt Independence Criterion* (HSIC) [Gretton et al., 2005, Camps-Valls et al., 2010]. Such a non-parametric measure can be computed thanks to kernel matrices $\boldsymbol{K}'$ and $\boldsymbol{K}'_{YY}$ (of size $n \times n$), corresponding to X' and Y', respectively:

$$\mathrm{HSIC}(X', Y') = (1/(n-1)^2)\mathrm{Tr}(\boldsymbol{H}\boldsymbol{K}'\boldsymbol{H}\boldsymbol{K}'_{YY}) \, . \tag{7.6}$$

After substituting the kernel matrix $\boldsymbol{K}^*$ representing the projections (replacing $\boldsymbol{K}'$) and the matrix based on the labels $\boldsymbol{K}^*_{YY}$ (replacing $\boldsymbol{K}'_{YY}$) into the HSIC formula (7.6), and dropping the unnecessary scaling factor, our objective becomes the maximization of

$$\mathrm{Tr}(\boldsymbol{H}(\boldsymbol{K}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{K})\boldsymbol{H}\boldsymbol{K}^*_{YY}) = \mathrm{Tr}(\boldsymbol{W}^\top\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^*_{YY}\boldsymbol{H}\boldsymbol{K}\boldsymbol{W}) \, . \tag{7.7}$$

Locality preservation

The second purpose, the locality preservation, is attained through a manifold regularizer enforcing smoothness with respect to the underlying data geometry, i. e. a regularizer for which small variations over the manifold lead to small variations in the projection [Belkin et al., 2006]. To this end, we first build the graph Laplacian matrix $\boldsymbol{\mathcal{L}} = \mathbb{D} - \boldsymbol{M}$, where $\boldsymbol{M}$ is an *affinity matrix* of elements $M_{i,j} = \exp(-d^2_{i,j}/2\sigma^2)$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are k-nearest neighbors (with an Euclidean distance $d_{i,j}$ in the input space) and $M_{i,j} = 0$ otherwise. $\mathbb{D}$ is a diagonal *degree matrix* with elements $\mathbb{D}_{i,i} = \sum_{j=1}^{n_S+n_T} M_{i,j}$. Since we would like that samples that were close in the initial data space remain close in the transformed space, the goal is to minimize

$$\frac{1}{(n_S+n_T)^2}\sum_{i,j} M_{i,j} \left\| \boldsymbol{x}^*_i - \boldsymbol{x}^*_j \right\|^2 = \frac{1}{(n_S+n_T)^2}\mathrm{Tr}(\boldsymbol{W}^\top\boldsymbol{K}\boldsymbol{\mathcal{L}}\boldsymbol{K}\boldsymbol{W}) \, , \tag{7.8}$$

where $\boldsymbol{x}^*_i$ and $\boldsymbol{x}^*_j$ are the projections of the initial samples.

Ultimately, including the manifold regularization term, we may formulate the final optimization problem of SSTCA as

SSTCA kernel learning problem

$$\begin{aligned} \arg\min_{\boldsymbol{W}} \quad & \Big\{ \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{L}\boldsymbol{K}\boldsymbol{W}) + \mu \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{W}) \\ & + \frac{\lambda}{(n_S + n_T)^2} \mathrm{Tr}(\boldsymbol{W}^\top \boldsymbol{K}\mathcal{L}\boldsymbol{K}\boldsymbol{W}) \Big\} \\ \text{s.t.} \quad & \boldsymbol{W}^\top \boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^*_{YY}\boldsymbol{H}\boldsymbol{K}\boldsymbol{W} = \boldsymbol{I}\,, \end{aligned} \tag{7.9}$$

where $\lambda \geq 0$ is a tradeoff parameter weighting the importance of the local manifold. Note that in the following, for simplicity's sake, $\frac{\lambda}{(n_S+n_T)^2}$ is directly referred to as λ. As for unsupervised TCA, the optimization can be reformulated as a trace maximization problem solved via the following generalized eigenvalue problem:

$$\boldsymbol{K}\boldsymbol{H}\boldsymbol{K}^*_{YY}\boldsymbol{H}\boldsymbol{K}\boldsymbol{v} = \rho(\boldsymbol{K}(\boldsymbol{L} + \lambda\mathcal{L})\boldsymbol{K} + \mu\boldsymbol{I})\boldsymbol{v}\,. \tag{7.10}$$

The projection matrix $\boldsymbol{W}$ and the subsequently derived m SSTCA transfer components for some new samples are obtained in the same fashion as for TCA.

7.4 DATA AND EXPERIMENTAL SETUP

For our analyses we utilized two different datasets bearing different degrees of shift between source and target domains.

7.4.1 *Hyperspectral ROSIS image of Pavia*

Image and ground truth

The first dataset is the hyperspectral ROSIS image of Pavia (Italy) presented in Appendix B.3 (see page 156). For our adaptive thematic classification task, we took into account the 4 classes appearing throughout the entire scene: "buildings", "roads", "shadows" and "vegetation". Details are available in Tab. B.3.

Sample selection bias: light dataset shift

As one can see from Fig. B.3, the spatial extent of the source sub-image is quite small, involving a description of the classes that is presumably not rich enough to account for the complete variation of the spectral signatures over the entire image. This raises the question of the representativeness of the training samples for classification. Hence, adaptation is required to correct a sample selection bias problem. The dataset shift level is qualified as "light", as it is illustrated by the scatterplots of the top row of Fig. 7.2.

7.4.2 *VHR QuickBird images of Zurich*

Images, pre-processing and ground truth

The second dataset used here is the same already used in Chapter 6: we utilized the two VHR QuickBird images of Zurich (see Appendix B.1). To enhance the spatial information content of the scenes, the standard set of QuickBird bands has been extended by the same textural and morphological

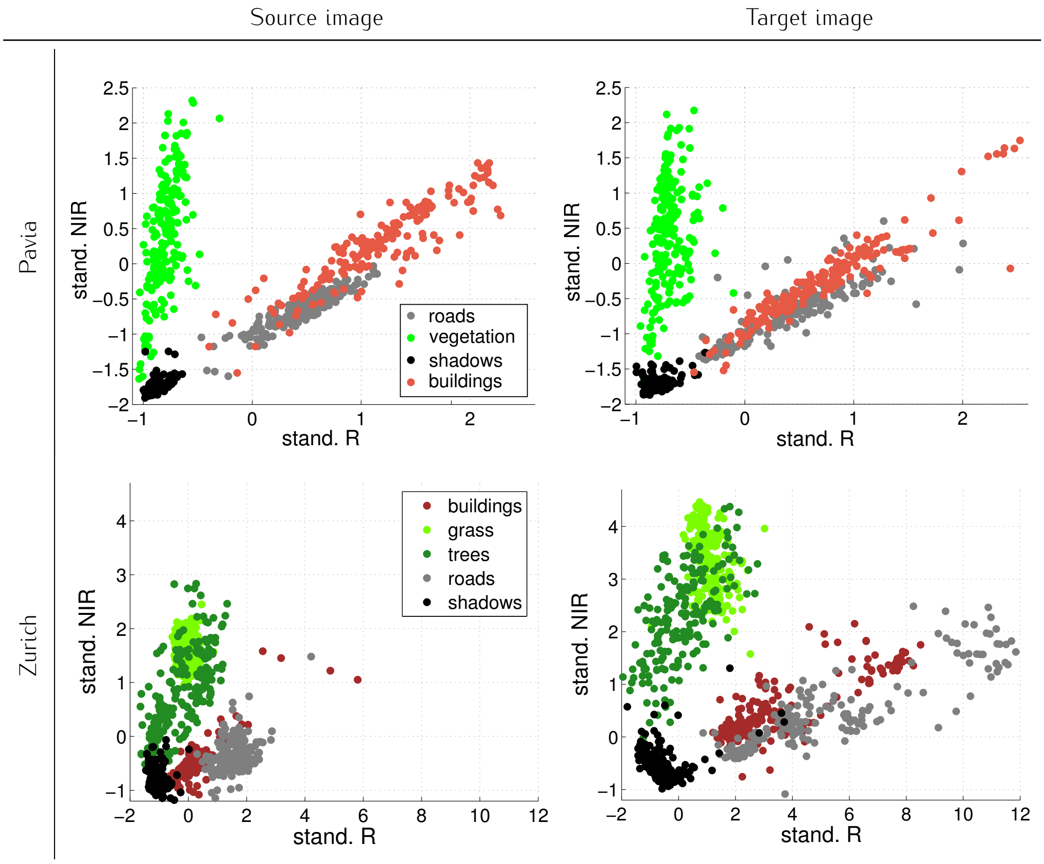


Figure 7.2: Raw DN data (standardized variables) red (R) vs. near-infrared (NIR) scatterplots of the (left) source and (right) target images of the (top) Pavia dataset with ROSIS band #49 vs. #95 and (bottom) Zurich dataset with QuickBird band #3 vs. #4.

features described in Section 6.3.1. Additionally, a textural feature based on correlation has been added to favor even more the label smoothness in the spatial domain. Thus, we obtained a final set of 16 variables: 4 VNIR bands, 1 PAN band, 5 textural features and 6 morphological features. For the sake of comparison, results obtained exclusively using the 5 spectral bands are reported in Section 7.5.6. For the classification task in this urban setting, again, we considered the 5 classes shared by the two images listed in Tab. B.1.

Moderate dataset shift

As mentioned, between the two acquisitions we notice strong differences in illumination conditions induced by changes in the sun elevation and in the acquisition geometry. Moreover, seasonal effects affecting the vegetation and the different nature of the materials used for roofs and roads increase the differences between images. This shift alters both the marginal and the class-conditional probabilities governing the two images: the deviation is thus judged as "moderate" in this case (see bottom row of Fig. 7.2).

7.4.3 *Experimental setup*

Proposed and competing FE methods

In the experiments below we compared TCA and SSTCA to the following other FE techniques: the classical PCA (see Section 3.5.1), its non-linear kernel-based counterpart, the KPCA (see Section 3.5.2), and the supervised kernel-based GDA mentioned in Section 3.4.2.

Image sampling strategies

With respect to the sampling strategies outlined in Section 7.2, TCA and SSTCA were always applied on a set X extracted from both the source and target domains, $X \subseteq X_S \cup X_T$, as a joint extraction was required by definition (strategies simply named **TCA** and **SSTCA**). Regarding the competing methods, PCA and KPCA have been applied considering a FE based on a single domain, the source image, i. e. $X \subseteq X_S$ (see Section 7.5.5 for a comparison to the $X \subseteq X_S \cup X_T$ setting). We termed these strategies **PCA_1DOM** and **KPCA_1DOM**, respectively. The application of GDA obligatorily requires label information, so the only setting adoptable was $X \subseteq X_S$ (**GDA_1DOM**).

Selection of the hyper-parameters

For the two TCA methods, previous works [Pan et al., 2011] suggested as valid the standard value of 1 for the tradeoff parameter μ. Likewise, for SSTCA we fixed the label dependence parameter $\gamma = 0.5$ equally balancing label dependence and orthogonality, whereas we varied the more critical locality parameter λ in $\{0, 10^{-4}, \dots, 10^{4}\}$ and the number of neighbors k in $\{10, 50, 100, 150, 200\}$. For the four kernel-based techniques (KPCA, GDA, TCA and SSTCA), we employed Gaussian RBF kernels, with the σ parameter selected as the median Euclidean distance among the data points used to define the mapping. We adopted this approach also for the selection of the scaling parameter σ needed in the computation of matrix $\boldsymbol{M}$ in SSTCA.

Preprocessing, datasets sizes & dimensions

For both the Pavia and Zurich datasets, all the initial features have been standardized to zero mean and unit variance, based on the descriptive statistics of the source domain. Sets of 200 pixels per class were used to define the projections (similar behaviors have been observed with sets of 150 and 300 pixels). The same samples constituted the training sets for the thematic classification. When pixels X_T were needed, $200 \cdot c$ unlabeled target pixels were randomly selected in the corresponding image. A total of 10 independent realizations of these sets has been used for the experiments to ensure a robust comparison. After exploratory analyses, the maximum number of extracted features was fixed to 18 and 15 for the Pavia and Zurich datasets, respectively.

Classifiers and baselines

To assess the suitability of the proposed FE methods, after the cardinal projection step, we proceeded with a classification by applying two intrinsically different classifiers trained on the transformed source training samples X_S^*. We made use of a parametric classifier, the LDA model describing all the classes by a common covariance matrix (see Section 3.4.1), and a kernel-based non-parametric classifier, the soft margin linear SVM with a penalty parameter C tuned in the range $\{10^{-1}, \dots, 10^{4}\}$ by 5-fold cross-validation. After FE, the classifiers have been trained with source samples mapped into a space of increasing dimension. With the LDA model we run the classifier for each number of dimensions. With the linear SVM we run the model for

Table 7.1: FE methods and baselines compared in the classification experiments (predictions on the target image).

Name	FE method	FE based on	Training on
`Tgt`	none	–	X_T
`Src`	none	–	X_S
`PCA_1DOM`	PCA	X_S	X_S^*
`KPCA_1DOM`	KPCA	X_S	X_S^*
`GDA_1DOM`	GDA	X_S	X_S^*
`TCA`	TCA	$X_S \cup X_T$	X_S^*
`SSTCA`	SSTCA	$X_S \cup X_T$	X_S^*

every third dimension starting at 2 extracted features for all FE techniques to reach convergence, except for GDA, for which we trained a SVM per additional feature. As reference upper and lower baselines, we also reported performances of models built using samples exclusively belonging to the target or source images (`Tgt` and `Src`). In these cases, the input space was constituted by the original spectral bands X_S or X_T (plus spatial information for the Zurich images). Table 7.1 summarizes the considered settings and the related names. The overall quality of the classification has been assessed by means of the Kappa statistic (see Appendix A).

HM scenarios

To carry out these classification tasks, we considered two scenarios:

1. FE methods applied to the raw data, without any initial matching of the source and target images.

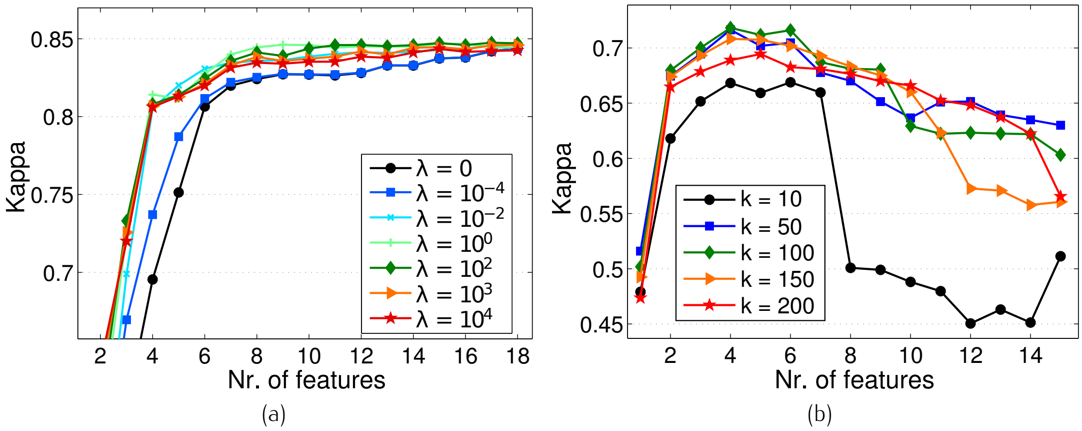
2. FE methods applied after a preprocessing with HM, taking the source image as reference image.

The accuracy assessment in the target domain has been done by means of an independent test set counting 14,047 pixels for the Pavia dataset and 26,797 samples for the Zurich dataset.

7.5 RESULTS

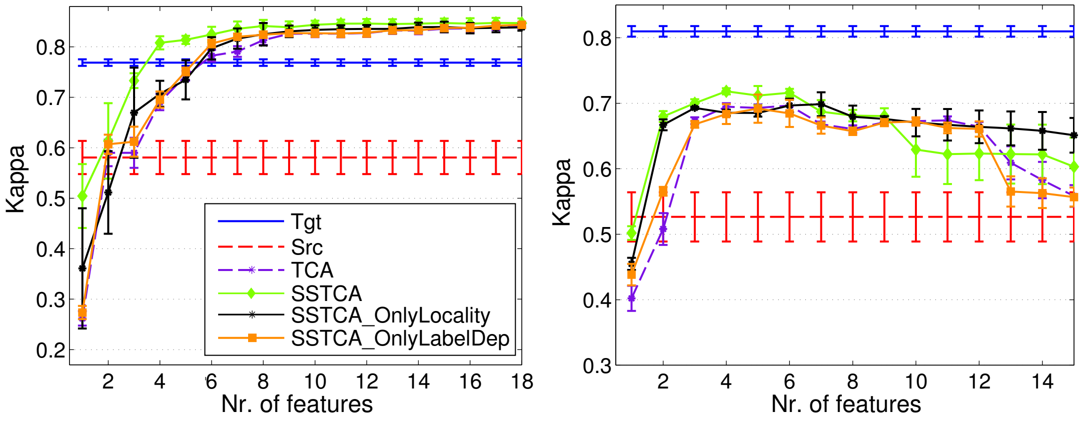
7.5.1 *Analysis of SSTCA parameters*

In Fig. 7.3, we report the results of a sensitivity analysis of SSTCA involving its two critical parameters: the λ parameter controlling the importance of the locality preserving term and the number of neighbors k used to build the graph Laplacian $\mathcal{L}$. We focus on data after HM and we employ a LDA classifier. As illustrated in Fig. 7.3(a) for the Pavia dataset, λ returned the best overall performances when set as $\geq 10^{-2}$. Concerning the number of neighbors, as shown in Fig. 7.3(b) for the Zurich images, k seemed to reach an optimum in classification accuracy if chosen between 50 and 150. Therefore, for the rest of the study, we set $\lambda = 10^2$ and $k = 100$.



(a) (b)

Figure 7.3: LDA classification performances on the target image (average of estimated Kappa statistic over 10 runs) after SSTCA ($\mu = 1$, $\gamma = 0.5$). (a) Behavior of the λ parameter on the Pavia dataset with k fixed to 100. (b) Behavior of the k parameter on the Zurich dataset with λ fixed to 10^2. Datasets after HM have been considered.



(a) (b)

Figure 7.4: LDA classification performances on the target image (average and standard deviation of estimated Kappa statistic over 10 runs) on the (a) Pavia and (b) Zurich datasets to assess the influence of locality preservation and label dependence in the SSTCA optimization problem. **SSTCA_OnlyLocality**: SSTCA run with the γ parameter controlling the label dependence set to 0. **SSTCA_OnlyLabelDep**: SSTCA run with the λ parameter controlling the locality preservation set to 0. Rest of the SSTCA parameters kept as described in Section 7.4.3. Datasets after HM have been considered. Legend of (a) is also valid for (b).

With Fig. 7.4 we depict the influence of the peculiar objectives of SSTCA (extending the goals of unsupervised TCA) by analyzing the impact of the associated terms appearing in the optimization problem: the locality preservation term and the label dependence term (see Section 7.3.2). In both Figs. 7.4(a) and 7.4(b), relating to the Pavia and Zurich images respectively (HM setting and LDA classifier also in this case), we first remark the overall best performance of the standard SSTCA-based extraction (`SSTCA`: solid light green line). Indeed, when both the terms enter the optimization procedure, the benefit of SSTCA over the unsupervised TCA (`TCA`: solid purple line) is tangible. On the one hand, if only the locality term influences the definition of the mapping (`SSTCA_OnlyLocality`: dashed black line), we observe classification accuracies superior to those obtained after TCA, especially in the case of the Zurich image. On the other hand, when optimizing the label dependence only (`SSTCA_OnlyLabelDep`: solid orange line), we do not remark any significant improvement over TCA. Therefore, this experiment confirms the usefulness of the combination of the two objectives within SSTCA.

7.5.2 *Classification performances*

Figure 7.5 illustrates the performances of the previously depicted DA strategies via the cross-domain classification accuracies achieved on the target image. The left-hand side of the figure refers to the Pavia dataset while the right-hand side reports on the Zurich experiments.

7.5.2.1 *Pavia ROSIS dataset*

Figure 7.5(a) shows the results obtained using LDA without HM, whereas Fig. 7.5(b) depicts the behavior of the same LDA classifier after HM. A significant gap is noticeable in both plots between source and target-based models. Nevertheless, the influence of HM as a preprocessing step is remarkable. Indeed, LDA classifiers trained on original target data (`Tgt`: solid blue line) outperform LDA models based on original source data (`Src`: dashed red line) by 0.356 Kappa points when no matching is performed, while this difference reduces to 0.188 Kappa points when applying HM.

Moreover, in Figs. 7.5(a) and 7.5(b), we remark three separate trends. First, a PCA-based FE (`PCA_1DOM`: dashed dark green line), reveals a performance just above the baseline of the `Src` model. Peak accuracies are observed in both experiments with 2 features, while as noisy features related to smaller eigenvalues are provided to the classifier, the quality of the model deteriorates (see Figs. 7.11 and 7.12 for a representation of these features).

Second, we note comparable evolutions for the kernel-based FE techniques TCA and KPCA (`TCA`: dashed purple line, `KPCA_1DOM`: solid brown line), yielding a robust and much more satisfactory performance reaching and even exceeding the accuracy of the target model. When using less than 9 (no HM) or 8 (with HM) extracted features, TCA seems to be the more reliable of the two methods, while as more features come into play they converge to similar performances. A possible reason behind the observed tendency is that the two feature extractors share some properties and objectives. Both are non-linear kernel methods aiming at the maximally preserving the variance of the original data.

Third, the behavior across the entire range of features of SSTCA (`SSTCA`: solid light green line) stands out for its remarkable accuracies being much

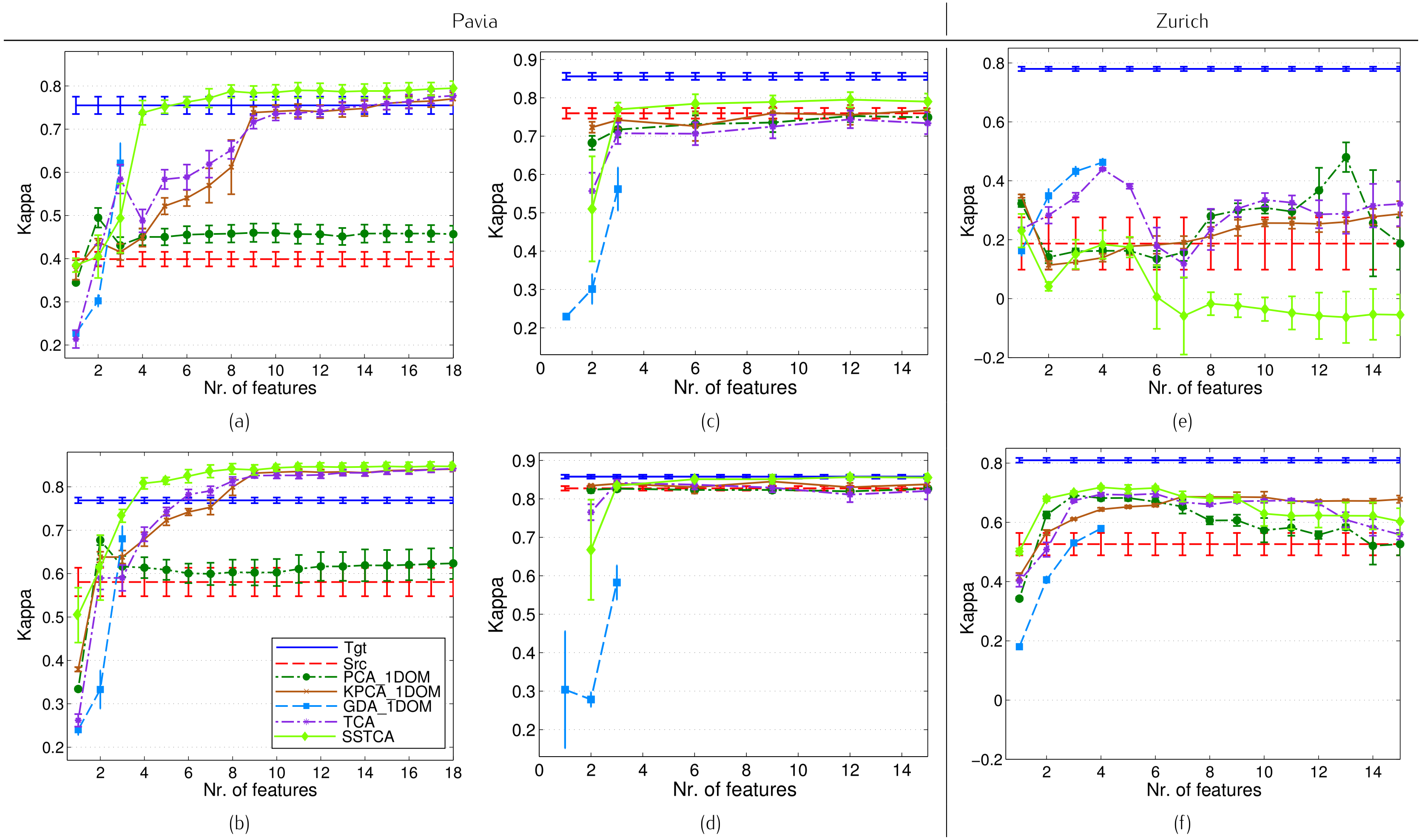


Figure 7.5: Classification performances on the target image (average and standard deviation of estimated Kappa statistic over 10 runs) on the (left) Pavia and (right) Zurich datasets considering different classifiers and settings. (a) LDA on the Pavia dataset without HM. (b) LDA on the Pavia dataset with HM. (c) Linear SVM on the Pavia dataset without HM. (d) Linear SVM on the Pavia dataset with HM. (e) LDA on the Zurich dataset without HM. (f) LDA on the Zurich dataset with HM. Legend of (b) is valid for all the panels.

higher than the rest of the FE methods from 4 (no HM) or 3 (with HM) features on. The performance with unmatched data (Fig. 7.5(a)) is particularly satisfactory since the associated Kappa statistic is within the error bar of the target model with just 4 features (compared to 9 or more for the other kernel-based methods).

Finally, the GDA technique (`GDA_1DOM`: dashed light blue line) provides reasonable accuracies competing with those of the other FE methods when extracting 3 features. However, since the maximum number of features it can extract is bounded by $c - 1$, this approach can not go beyond this level of precision.

Considering the classification with a linear SVM, Fig. 7.5(c) illustrates the results obtained without HM, while Fig. 7.5(d) refers to the HM case. We first remark the general improvement in the precision of the classification over LDA ($\approx$ 0.1 Kappa points increase of the upper reference accuracy provided by the `Tgt` approach). The tendency for the two baselines `Tgt` and `Src` is basically the same: a large reduction of the gap is ensured by a pre-processing with HM. However, thanks to the intrinsic properties of the SVM, we observe that the differences between the other curves are reduced. In both cases (with and without HM) the only technique able to positively differ from the rest is SSTCA. In Fig. 7.5(c), this semisupervised method allows to outperform (from 3 features on) the already very robust SVM model trained on the original source image (`Src`) while the other FE methods remain below this baseline. In Fig. 7.5(d), SSTCA again is the only strategy attaining the precision of the `Tgt` model. In both figures, we note the unsatisfactory performance by GDA, far below the rest of the strategies.

7.5.2.2 *Zurich QuickBird dataset*

Presented in the right-hand column of Fig. 7.5, the results referring to the Zurich dataset when using the LDA classifier allow us strengthen the conclusions drawn above. Indeed, Figs. 7.5(e) and 7.5(f) confirm the usefulness of HM, as all the methods we employed manifestly fail if applied to unmatched data. The reason behind this result, particularly marked for the SSTCA, is that, since the shift is larger than in the Pavia dataset, no well-behaved cross-domain relationships may be extracted. On these images where the dataset shift is moderate, the Kappa accuracy of the straightforward application of the original source LDA classifier (`Src`) on target image improved from 0.187 to 0.527 after HM. The HM preprocessing allowed the FE models to correctly find the principal directions of data variation that are shared across domains.

Analyzing more precisely Fig. 7.5(f), we observe a similar behavior as for the Pavia dataset. A larger number of features are needed by the TCA and KPCA approaches to achieve good performances with respect to PCA (best performance with 3 features). Nonetheless, these accuracies are still more than 0.1 Kappa points below the same-domain target model (`Tgt`). In this situation as well, the SSTCA mapping proves to be the most appropriate (best Kappa = 0.718, obtained with 4 features), at least until the inclusion of the 7th feature.

On the contrary, the GDA-based FE performs poorly, with only a minimal improvement over the `Src` baseline. This poor performance, already noted on the Pavia dataset, suggests that only focusing on the maximization of the class separation in the source domain (the sole image where class labels are known) is detrimental when the purpose is to find a shared data space in which domain invariance too has to be ensured.

The performance of unsupervised TCA being systematically exceeded by SSTCA indicates that the reduction of the statistical deviation between datasets alone (as measured by the MMD) is not sufficient to achieve a good portability of the classifiers across domains. Indeed, we stress the importance of jointly considering the label dependence and the locality preservation objectives in the definition of the mapping function. When these components are appropriately combined, as done by SSTCA, the MMD minimization goal guarantees a suitable knowledge transfer among the images.

Additionally, we remark that exclusively pursuing a better class discrimination, especially since the latter is sought based on class labels from the source domain only, excessively deforms the input space. Such a drawback harms the FE via GDA, preventing thus a proper classifier adaptation. SSTCA ensures a better tradeoff, as it simultaneously considers geometric preservation and class discrimination.

In general, another main finding consists in the complementarity of the two key pre-classification procedures: HM and FE. On both sets of images, the best accuracies were reached by models built on images with matched histograms having undergone the FE. After these processing steps, the source and target datasets are sufficiently aligned and the features are discriminant enough to allow classifiers trained on one image to generalize well on the other too.

7.5.3 *Classification maps and individual class accuracies*

Figure 7.6 provides the LDA classification maps referring to the Pavia dataset (experiment with HM, run #1). Compared to the map produced by the target-based model `Tgt` (Fig. 7.6(a)), the thematic map obtained by the straightforward `Src` strategy is poorer, with an inadequate delineation of the built areas (Fig. 7.6(b)). A more acceptable result is obtained with a FE by PCA (Fig. 7.6(c)), but in this case many false alarms for the class "shadows" are observed. Changing the representation of our data in a non-linear fashion, as done by SSTCA (Fig. 7.6(d)), facilitates the LDA model in providing a precise thematic map. No recurrent errors appear and the resulting map seems less affected by noise.

Figure 7.7 reports the LDA thematic maps for the Zurich dataset (experiment with HM, run #1). The reference map, the one where the model has been trained on the target image (`Tgt`), is shown in Fig. 7.7(a). The `Src` approach (Fig. 7.7(b)) yields a map having a much lower accuracy. The most

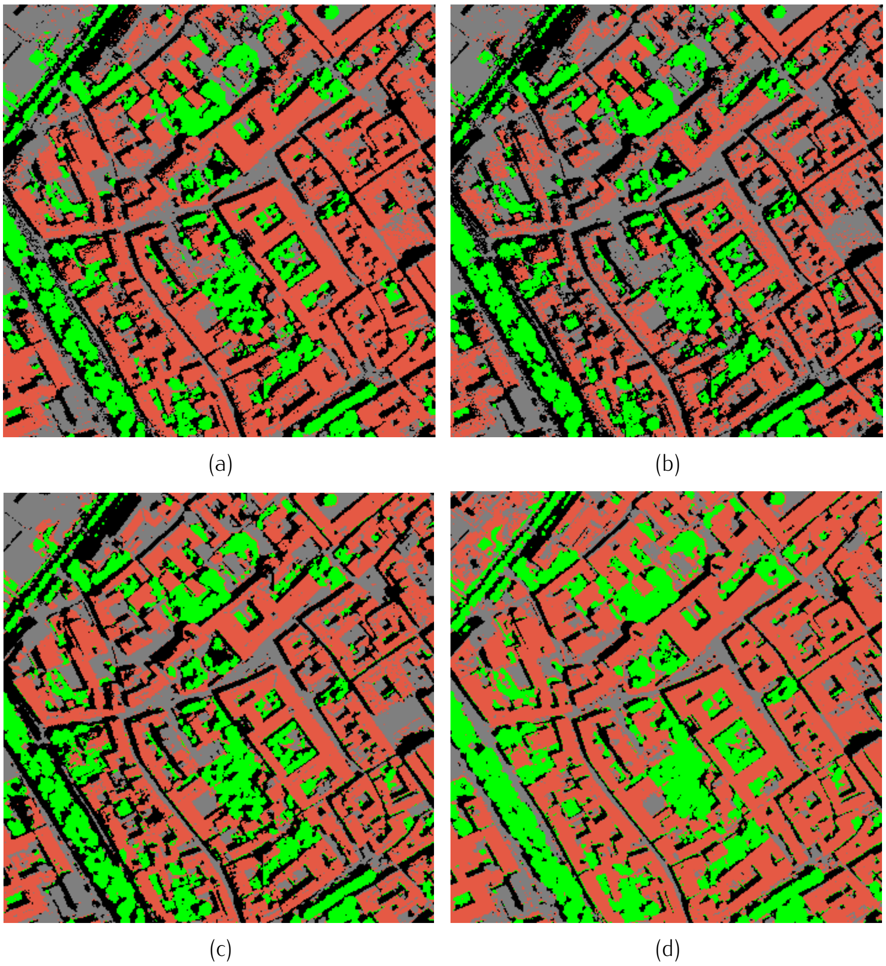


Figure 7.6: LDA classification maps on the Pavia target image, after HM (run #1). Legend: "buildings" → brown, "roads" → grey, "shadows" → **black**, "vegetation" → green. (a) `Tgt`: Kappa = 0.769. (b) `Src`: Kappa = 0.601. (c) `PCA_1DOM` (2 features): Kappa = 0.690. (d) `SSTCA` (11 features): Kappa = 0.864.

striking type of error relates to the almost complete failure in detecting the class "trees". On the contrary, few misclassifications are committed by the procedures involving a FE step beforehand, thus highly enhancing the overall quality of the map. Indeed, after a FE by PCA (Fig. 7.7(c)) or by SSTCA (Fig. 7.7(d)), the two vegetation classes are fairly correctly detected while keeping a good characterization of the buildings. However, in both cases, false alarms concerning this class, with a negative impact mostly on the class "roads", appear throughout the map.

Let us now focus on the evolution of the LDA individual class accuracies, assessed via the F-measure (see Appendix A), as a function of the number of extracted features both for the Pavia (Fig. 7.8) and Zurich (Fig. 7.9) datasets. On the Pavia dataset, we first analyze class-specific trends of the `PCA_1DOM` method thanks to Fig. 7.8(a). We can identify a distinct F-measure peak with 2 features for the classes "roads" and "buildings". It is

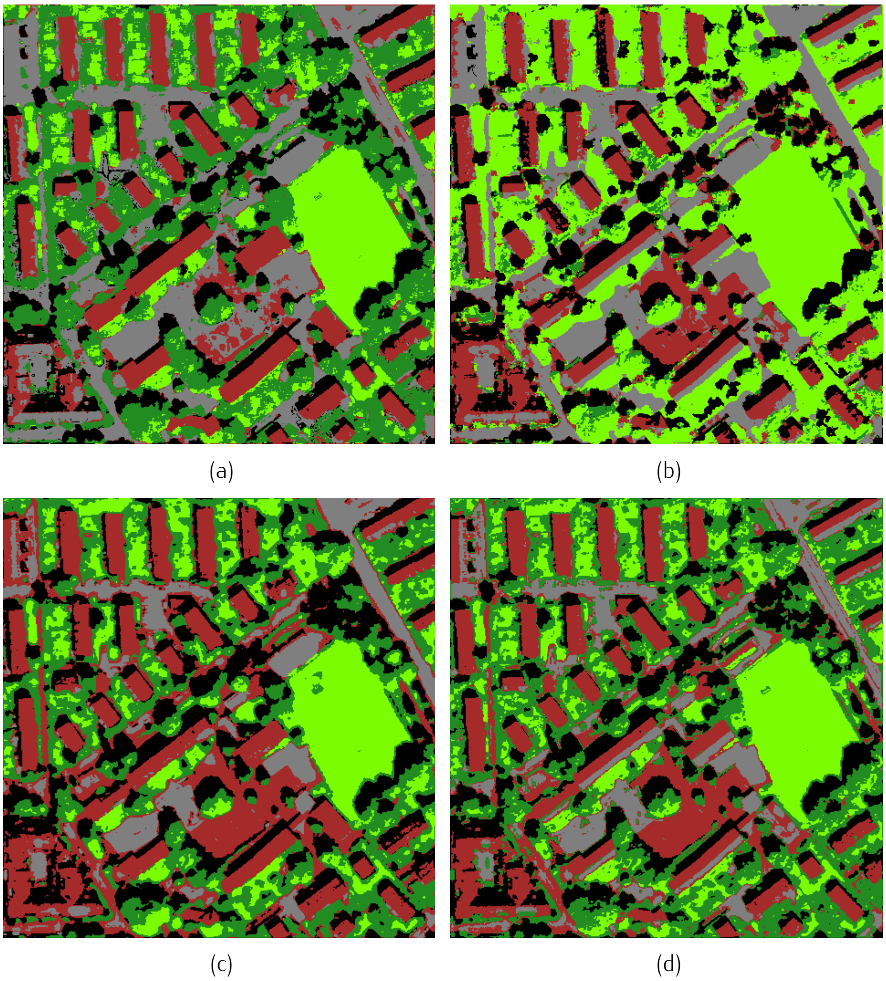


Figure 7.7: LDA classification maps on the Zurich target image, after HM (run #1). Legend: "buildings" → brown, "roads" → grey, "shadows" → **black**, "trees" → dark green, "grass" → light green. (a) `Tgt`: Kappa = 0.806. (b) `Src`: Kappa = 0.573. (c) `PCA_1DOM` (3 features): Kappa = 0.694. (d) `SSTCA` (4 features): Kappa = 0.716.

thus straightforward to explain the maximum in Kappa statistic observed in Fig. 7.5(b) for the PCA-based system with the appropriate detection of these two types of land-cover. The performance of the `SSTCA` technique is illustrated in Fig. 7.8(b) instead. Aside from a general increase in all the per-class accuracies, the most apparent advantage of SSTCA over PCA lies in the correct classification of the class "shadows", a category which was highly mishandled by the linear feature extractor.

Turning to the Zurich dataset, Fig. 7.9(a) illustrates the behavior of the `PCA_1DOM` approach. We notice the good precisions for all classes in the 3 to 5 features region, leading to the peak in overall accuracy visible in Fig. 7.5(f). This class specific plot allows us to have a good insight on the reasons behind the Kappa statistic decrease from feature #6 on. It is indeed the class "trees" that shows a decreasing trend in precision as more

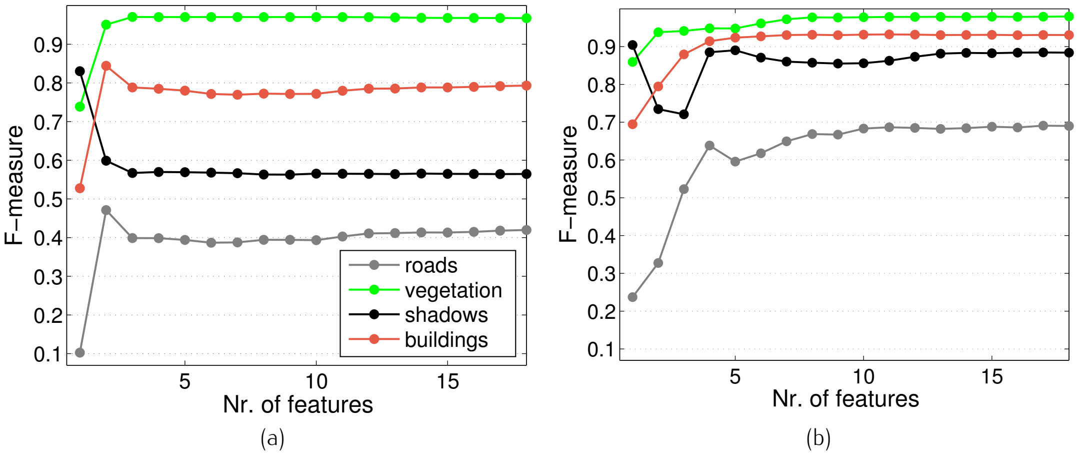


Figure 7.8: LDA individual class accuracies (average of F-measure over 10 runs) on the Pavia target image after HM. (a) **PCA_1DOM**. (b) **SSTCA**. Legend of (a) is also valid for (b).

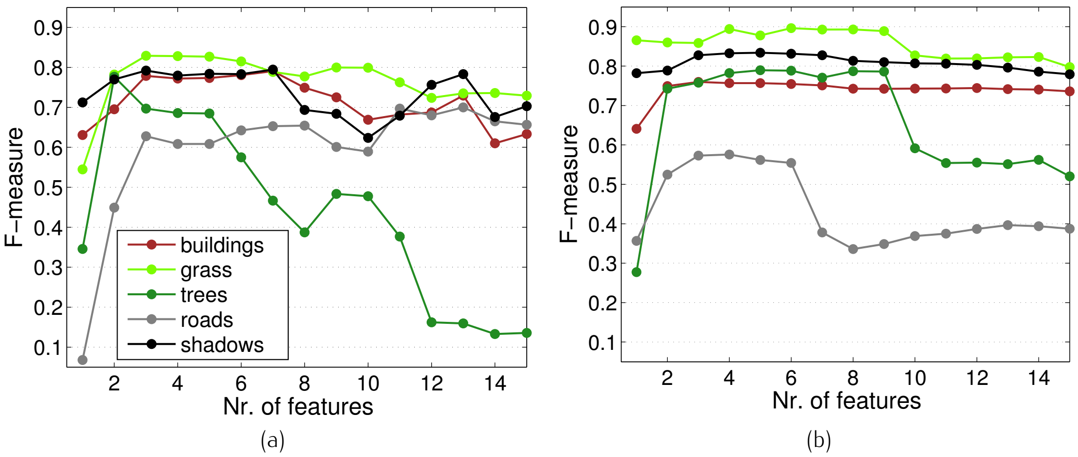


Figure 7.9: LDA individual class accuracies (average of F-measure over 10 runs) on the Zurich target image after HM. (a) **PCA_1DOM**. (b) **SSTCA**. Legend of (a) is also valid for (b).

features are extracted by PCA, probably indicating that these components possessing a high spatial frequency badly affect the discrimination of this class. Figure 7.9(b) breaks down the performance of the **SSTCA** method and reveals that the critical land-cover is in this case the class “roads”, whose F-measure curve evolves far below the rest. Since the “buildings” and “roads” thematic classes are spectrally very similar, the label dependency term cannot convey information useful for discrimination, and a naïve approach may result in better class-specific scores. This confirms the visual inspection of the classification map in Fig. 7.7(d).

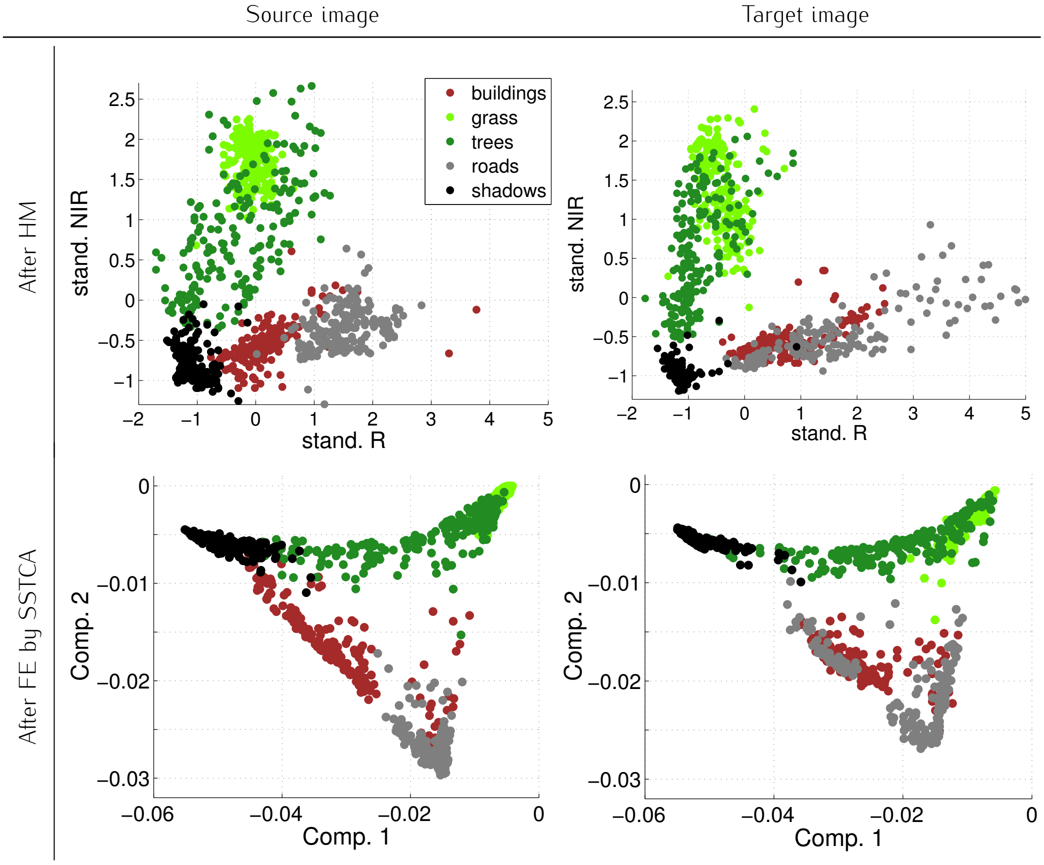


Figure 7.10: Scatterplots of the (left) source and (right) target images of the Zurich dataset (run #1). (top) Data after HM in the red (R) vs. near-infrared (NIR) space (standardized variables). (bottom) Data after HM and after FE by SSTCA based on $X \subseteq X_S \cup X_T$ plotted in the 1st vs. 2nd component space.

7.5.4 *Visual analysis of the extracted features*

Referring to the Zurich QuickBird dataset, the scatterplots of Fig. 7.10 allow us to visually perceive the reduction of the shift (invariance property) between data distributions. Looking at the plots for the raw images (bottom row of Fig. 7.2 on page 92), we easily understand why a model trained in the source domain fails if applied on the target domain. The scatterplot of the classes of the source image is compressed towards the origin of the space since, for the vegetation class for instance, the reflected energy is lower due to the leaf senescence process associated with the season (image taken in autumn). After HM (top row of Fig. 7.10) the situation improves, with point clouds roughly occupying the same regions of the red vs. NIR space in the source and in the target domain. Nonetheless, overlapping classes and shifts in the class-conditional distributions explain the poor target classification maps produced by a model trained on the original, though histogram matched, source image (see Fig. 7.7(b)). If we apply the crucial FE step, via SSTCA for instance (bottom row of Fig. 7.10), we notice a clear decrease both in the general shift of the images and in the change in class distributions. In the space formed by the 1st and 2nd SSTCA components, the boundaries of the land-cover classes a supervised classification model learns on the source image are directly transferable to the target image, resulting in accurate thematic maps (see Figs. 7.5(f) and 7.7(d)). As noticed above, we may also observe the mixing of the "buildings" and "roads" classes.

Figures 7.11 and 7.12, both concerning the Pavia dataset, reveal the enhancement of class separability (discrimination property) induced by feature-representation-transfer methods. On the one hand, Fig. 7.11 illustrates the scatterplots after FE for source and target data in the space formed by the first two components (left) compared to those constituted by the 4th and 5th components (right). On the other hand, Fig. 7.12 visualizes the same information, but in the geographical space: the RGB compositions of the target image correspond to the 18 extracted components taken 3 at a time and in a decreasing variance order. Class discrimination in the first components is appropriately guaranteed by all the FE techniques illustrated here. In fact, well separated clusters (plus good superimposition of the source and target data) appear when looking at the 1st and 2nd features (Fig. 7.11) and clear structures are visible in the RGB image associated with the first 3 components (Fig. 7.12). As we investigate the subsequently extracted features, we remark some differences among the methods. In fact, PCA starts yielding noisy variables as soon as from the 4th derived feature onwards, resulting in a single large cluster mixing all the classes in the scatterplot and a noisy image in the RGB composition. This is the reason of the decrease in classification accuracy for the PCA-based system observable in Fig. 7.5(b) after the peak at 2 features. We appreciate the superior quality of kernel-based extractors (KPCA, TCA and SSTCA), almost always returning informative and discriminant features throughout the investigated range. Especially, with the plot of the 1st vs. 2nd component (4th row, left column of Fig. 7.11), we draw the attention to the ability of the SSTCA method in yielding a new space of features in which the classes are very well separated as a result of the label dependency goal pursued by the projection.

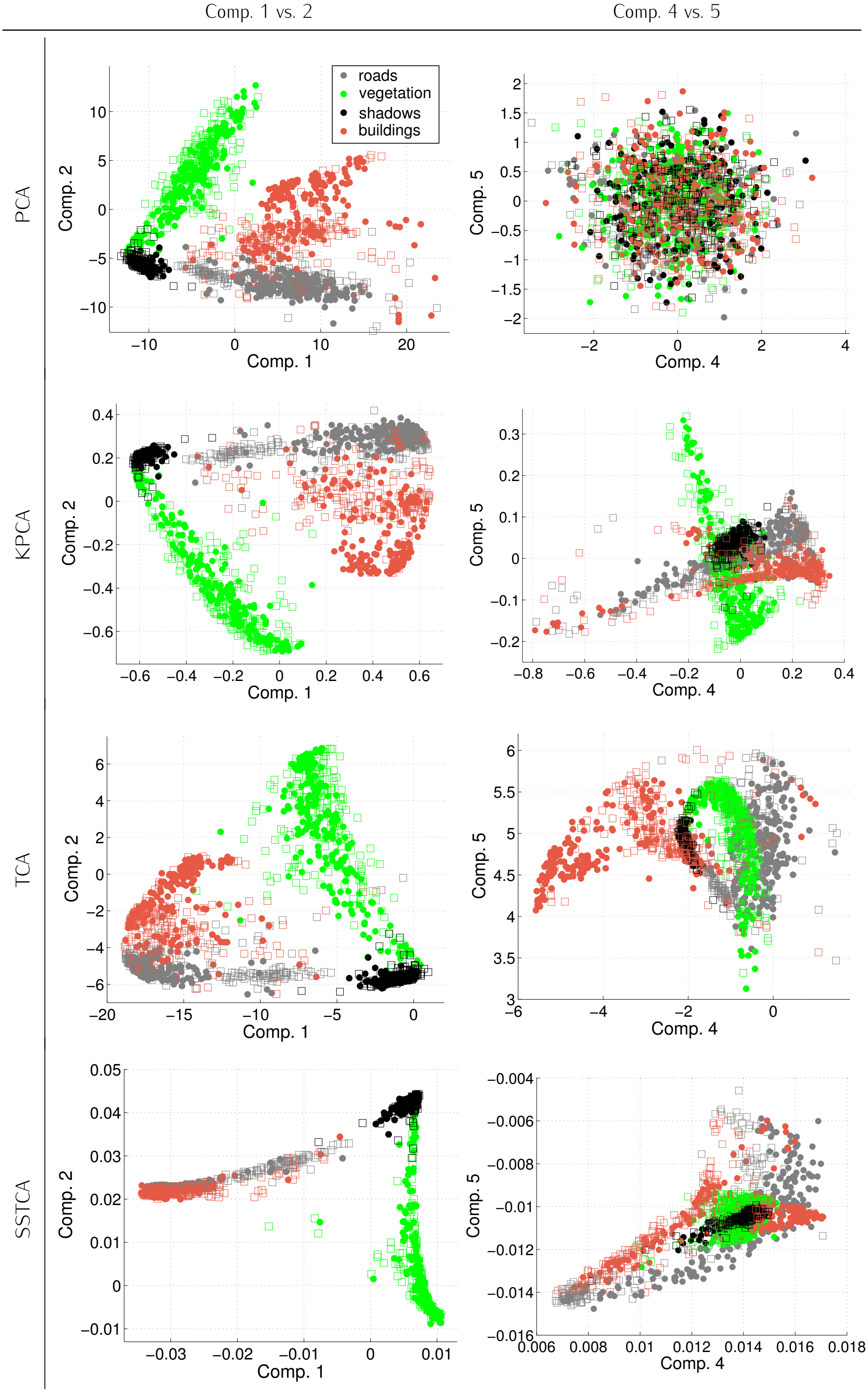


Figure 7.11: Scatterplots after FE in the (left) 1st vs. 2nd component and (right) 4th vs. 5th component space for the Pavia dataset (experiment with HM, run #1). Source data: ●, target data: □. (1st row) PCA based on $X \subseteq X_S$. (2nd row) KPCA based on $X \subseteq X_S$. (3rd row) TCA based on $X \subseteq X_S \cup X_T$. (4th row) SSTCA based on $X \subseteq X_S \cup X_T$.

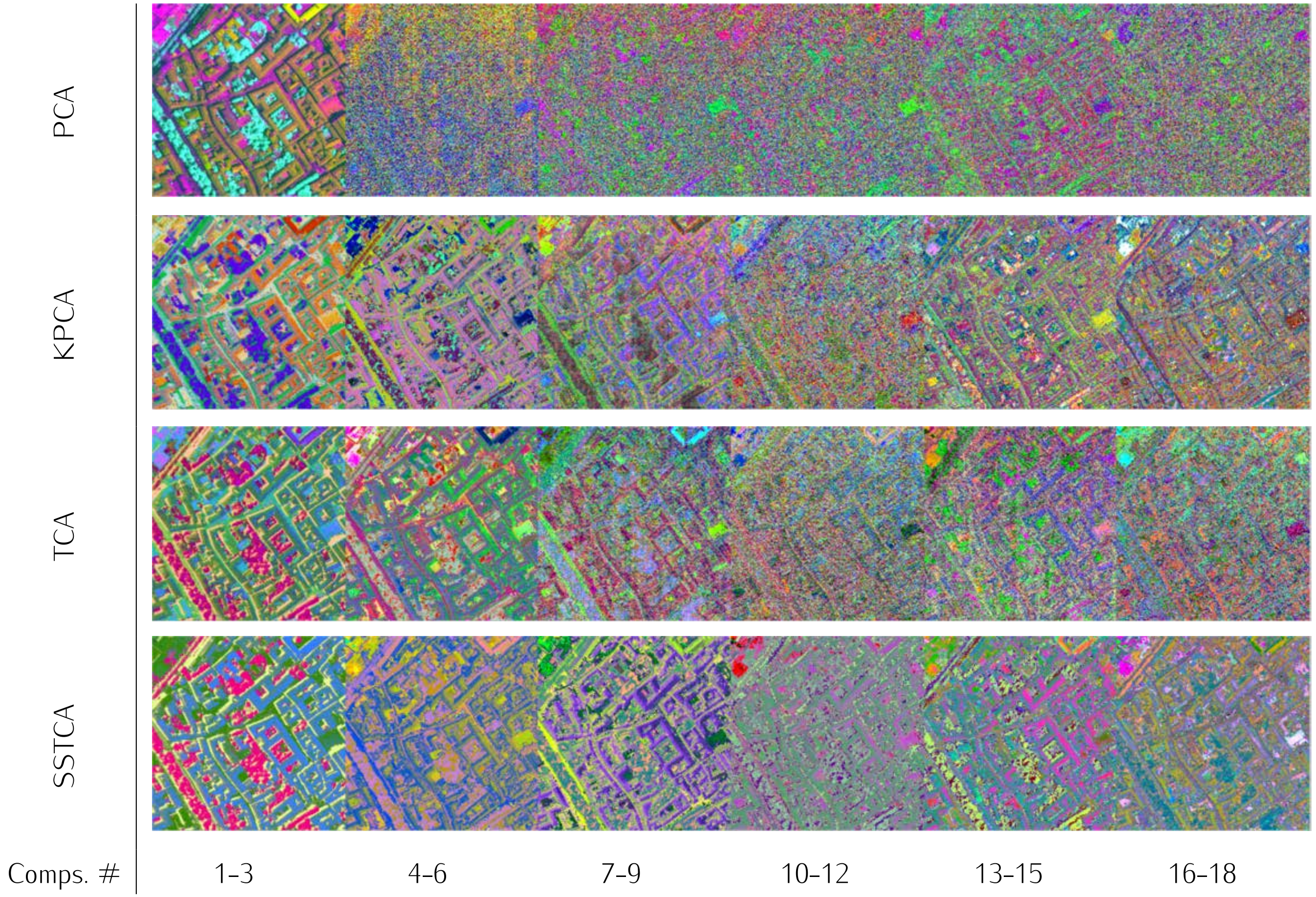
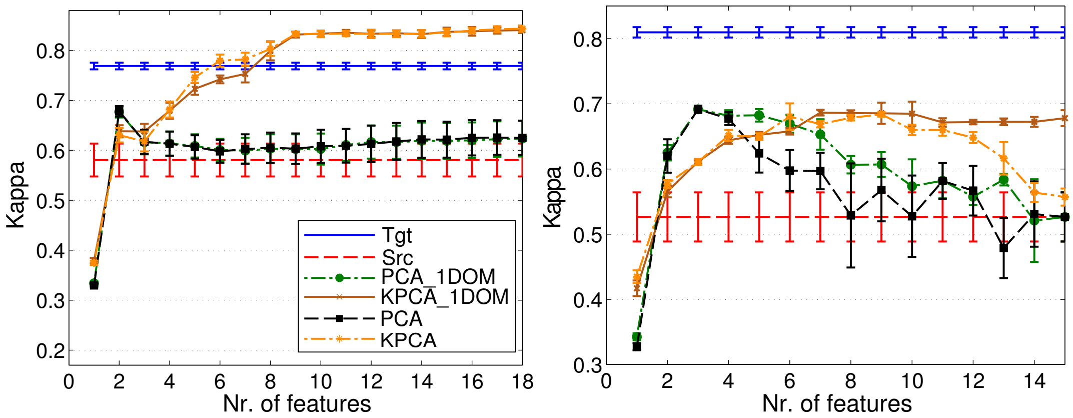


Figure 7.12: Visualization of the first 18 components extracted by the FE methods for the Pavia target image (experiment with HM, run #1). The columns represent the RGB combinations (3 components at a time) of the features sorted by decreasing eigenvalue. (1st row) PCA based on $X \subseteq X_S$. (2nd row) KPCA based on $X \subseteq X_S$. (3rd row) TCA based on $X \subseteq X_S \cup X_T$. (4th row) SSTCA based on $X \subseteq X_S \cup X_T$.



(a)

(b)

Figure 7.13: LDA classification performances on the target image (average and standard deviation of estimated Kappa statistic over 10 runs) on the (a) Pavia and (b) Zurich target images to test the sampling settings. Datasets after HM have been considered. Legend of (a) is also valid for (b).

7.5.5 *Influence of the origin of the samples*

We present here the results of an experiment aimed at gauging the influence of the domain of origin of the samples used to design the mapping function. To this end, for PCA and KPCA, the only methods allowing such a flexibility, we tested the sampling schemes described in Section 7.2. We compared the joint domain extraction based on $X \subseteq X_S \cup X_T$ (approaches named **PCA** and **KPCA**) to the single domain scheme involving pixels drawn from the source image only, i. e. $X \subseteq X_S$ (**PCA_1DOM** and **KPCA_1DOM**). Figure 7.13, referring to datasets after HM, reports the results of this experiment relying on LDA as the base classifier. On the Pavia image (Fig. 7.13(a)), we draw the attention to the almost indiscernible behavior between models built after the FE based on both domains (**PCA** and **KPCA** curves) and those based on the source domain only (**PCA_1DOM** and **KPCA_1DOM** curves). On the Zurich dataset (Fig. 7.13(b)), the single-domain FE settings even outperform the joint-domains counterparts for a wide range of number of features. Such a behavior suggests that using one domain only (the source image) as foundation for the FE does not imply a loss in invariance across domains. Instead, the overall trend in Kappa statistic of the classification system built after a single-domain FE can be judged superior to the FE involving two domains.

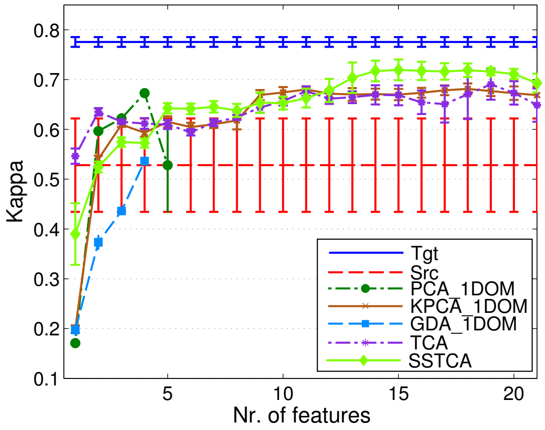


Figure 7.14: LDA classification performances on the target image (average and standard deviation of estimated Kappa statistic over 10 runs) on the Zurich target image using spectral bands only (4 VNIR + 1 PAN). Datasets after HM have been considered.

7.5.6 *Adaptation using exclusively the spectral bands*

In this Section, we briefly report on a test we carried out to gauge which is the real benefit of using the spatial information in the experiments on the Zurich data. Figure 7.14 presents the classification performances of the LDA classifier in a situation where both the FE and classification steps have been applied by considering only spectral data, i. e. the 4 VNIR bands and the PAN channel of the QuickBird acquisitions. In this case, with KPCA, TCA and SSTCA we extracted up to 21 features whereas we were limited to 5 components with PCA (bounded by the number of input variables d) and to 4 components with GDA (bounded by the number of classes $c - 1$). The results can be directly compared to those of Fig. 7.5(f) concerning the standard setting adopted in this Chapter for the Zurich data: a combination of spectral and spatial features.

Firstly, we notice the benefits of resorting to the spatial information for the single-image classification case, as testified by the purely spectral **Tgt** model only achieving an average Kappa of 0.775 compared to a Kappa of 0.810 when including the spatial features. Also, we observe the much larger variance in the results of the **Src** classifier with respect to that visible in Fig. 7.5(f), an indicator of the stability in the cross-image classification performance ensured by textural and morphological features. Second, a slight decrease in the performance of the **PCA_1DOM** and **GDA_1DOM** methods is detected for every size of the set of input features we tested. Third, the tendency for the **KPCA_1DOM**, **TCA** and **SSTCA** methods is to peak at the same level of accuracy as the previous spectral-spatial setting but by necessitating more features (≥ 15). Summing up, we can state that introducing contextual information in the DA exercise generally helps in augmenting the invariance of the input space used to describe the samples across domains.

7.6 CONCLUSIONS

Main achievements

In this Chapter, we studied the suitability of non-linear aligning transformations by FE in a remote sensing DA context. The purpose was to match the probability distributions of a target image to be classified to those of a source image whose labeled training data are already available. We verified the assumption that, after the proper projection, even a simple linear supervised classifier is able to accurately predict the land-cover across images.

Behavior of each of the FE methods

In this regard, we extensively analyzed the SSTCA technique by understanding the crucial role played by all the terms concurring in its optimization problem. The clear improvements over the unsupervised version, TCA, were apparent in all scenarios and regardless of the classifier, confirming thus the interest of a projection preserving the local geometry of the data as well as maximizing the dependence with the labels. Regarding the other considered FE techniques, a better performance of kernel-based methods such as KPCA and TCA with respect to simpler linear ones such as PCA has been distinctly noted. Nonetheless, these two kernel methods have demonstrated to perform almost equally throughout the entire set of scenarios we tested.

Usefulness of HM

Experiments also showed that the combination of HM with FE is extremely beneficial, pointing out the complementarity of these alignment strategies.

Discrimination and invariance

Furthermore, by means of dedicated visual representations of the derived features, we could provide important insights on the reasons behind the observed behaviors: class discrimination and domain-invariance are two fundamental properties the new features should display. SSTCA in particular has demonstrated to possess them both.

Benefits in concrete applications

To sum up, a well designed FE step will greatly improve the accuracy of a classifier trained on a different image. These findings represent a step ahead in defining effective tools for addressing large-scale land-cover mapping applications involving multiple remotely sensed images.

Further research

As regards the outlook on new research directions, a development we envisage is to broaden the set of FE techniques to be compared. In fact, especially among supervised methods making use of class labels, there is room for improvement in the definition a transformation that increases the class separability in both domains while preserving the data structure. Synergies between supervised FE and manifold alignment could prove highly beneficial. In addition, to further overcome the dataset shift, the application of FE strategies could be exploited in combination with classification techniques specifically designed for DA, the parameter-transfer approaches we reviewed in Sections 4.4.3 and 5.2.4.

8 ASSESSING ANGULAR DATASET SHIFT AND MODEL PORTABILITY IN MULTI-ANGLE IMAGE SEQUENCES

Outline: *This Chapter investigates the angular effects causing spectral distortions in multi-angle remote sensing imagery. We study two WorldView-2 multispectral in-track sequences acquired over urban areas. First, we quantify the degree of distortion affecting the sequences by means of the Maximum Mean Discrepancy. Second, we assess the ability of a classification model trained on an image acquired with a certain view angle to predict the land-cover of all the other images in the sequence. For both datasets, the efficacy of physically- and statistically-based normalization methods in obtaining angle-invariant data spaces is compared and synergies are discussed. In Section 8.1, we review the latest developments in exploiting multi-angle imagery as well as the main issues arising from a variable acquisition geometry. Subsequently, Section 8.2 introduces the multi-angle sequences we used. Section 8.3 presents the statistical quantification of the angular dataset shift. In Section 8.4, we tackle the question of the land-cover model portability. Section 8.5 concludes the Chapter and provides future research directions.*

8.1 INTRODUCTION

8.1.1 *Impact of the acquisition angle*

Acquisition angle & at-sensor radiance

The decrease of revisit time of satellites carrying VHR sensors has drastically increased the amount of data available to the end-users. These shorter collection intervals are, among other reasons, a consequence of the agility of the latest on-board control systems engineered to swiftly redirect the sensor even to very high off-nadir angles (see last part of Section 2.3.2). In this respect, as we mentioned in Section 2.5.2, it is acknowledged that the difference in the observation angle is a key factor altering the radiometry of remotely sensed images, together with atmospheric conditions, solar illumination, and phenology of vegetation [Schowengerdt, 2007]. Indeed, for every

This Chapter is part of a submitted paper that is now under review:

> G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia. Understanding angular effects in VHR in-track multi-angle image sequences and their consequences on urban land-cover model portability. *ISPRS Journal of Photogrammetry and Remote Sensing*, Submitted.

.

ground-cover type, the at-sensor radiance L_λ^s measured at the platform level (Eq. 2.1 on page 13) is dependent on the view angle. Three main angular physical phenomena are responsible for such a dependence.

Up-scattered path radiance

First, the optical depth of the atmosphere that the electromagnetic radiation has to go through before reaching the sensor has a critical impact on the acquisition. In essence, the lower the satellite elevation angle, the larger the off-nadir angle of acquisition, implying thus a longer optical path. In this situation, more radiance is scattered to the sensor by the atmosphere without any contact with the objects on the ground [Schott, 2007]. In Section 2.2, such a component of the total at-sensor radiance L_λ^s has been termed up-scattered path radiance and denoted with L_λ^{sp}. As we discussed, the distortion is driven by Rayleigh scattering, a physical phenomenon that is more marked at short wavelengths and at large off-nadir angles.

BRDF effects

Second, we observe BRDF effects, consisting in the variable scattering of an incident EM beam into the different directions of the hemisphere [Roujean et al., 1992, Schaepman-Strub et al., 2006]. The anisotropic scattering can be viewed as an angular property of each material. When imaging a given land-cover class, this causes brighter or darker surfaces depending on the satellite view angle with respect to the position of the illumination source.

Solar observational cross-section

Third, the solar observational cross-section, an effect responsible for changes in the reflectance of the objects with non-flat surfaces, is also considered a relevant factor. The consequences are clearly visible in the form of unmistakable differences in illumination and shadowing of parts of the surface (e. g. pitched roofs). This phenomenon is particularly apparent for objects with a considerable vertical structure. For instance, when the sensor acquires the image with a perspective similar to that of the illumination source (satellite and sun positioned at similar azimuth and elevation angles) for different types of trees the observed shadowing is minimized resulting in a brighter signal. A phenomenon known as *backward scattering hotspot* is distinctly detectable at these locations [Simmer and Gerstl, 1985, Hapke et al., 1996].

Shifted probability distributions

When the geometry of the acquisition varies from one image to another, all the phenomena above induce further shifts in the probability distribution of the spectra of the different land-cover classes that add to those considered up to now. Under these constraints, as discussed throughout this Thesis, the development of large-scale VHR land-cover/land-use mapping systems that require multiple remotely sensed images is hindered.

Recap of the available solutions

To overcome the shift in the image distributions and therefore to make classification routines more robust to angular effects (more portable across acquisitions) the two types of approaches we reviewed in Section 2.5.3 are usually adopted: absolute or relative normalization strategies. In the first category, we find atmospheric compensation approaches, whose purpose is to maintain the physical meaning of the image being processed, hence delivering final products describing the land-cover through quantities such as the surface reflectance $\rho(\lambda)$. The second category, instead, is mostly composed of statistical approaches aimed at adjusting the radiometry of a given image with respect to that observed on similar imagery.

8.1.2 *Exploiting and understanding the angular properties*

Stacking the angular images

The multi-angular capabilities of recently launched satellites have proven beneficial in many applications. Indeed, enriching the spectral information by simultaneously considering the series of angular images has often been instrumental in improving the thematic mapping or the information extraction over a particular scene. In urban studies, multi-angle sequences have been leveraged for detailed land-cover/land-use classification. This is due to the fact that urban materials such as asphalt and concrete possess distinct BRDF signatures [Puttonen et al., 2009]. Taking advantage of these properties, Duca and Del Frate [2008] provide encouraging results in the discrimination of urban structures using moderate resolution CHRIS data. Also exploiting CHRIS angular acquisitions, Verrelst et al. [2009, 2010] study the *Minnaert-k* parameter of the Rahman-Pinty-Verstraete model [Rahman et al., 1993] which allows to simulate BRDFs of various surfaces. The previously noted parameter describing the shape ("bowl" or "bell") of the reflectance curve in the angular domain is used to retrieve the density of the canopy cover in alpine forests. As regards WorldView-2, the potentialities in terms of urban classification offered by its VHR multi-angle sequences have been thoroughly analyzed in Longbotham et al. [2012a]. In order to include the angular reflectance profiles of the pixels in the classification problem, the authors provide a comprehensive investigation of various strategies going beyond the simple stacking of the images.

Model portability through the sequence

On the contrary, in Longbotham et al. [2012b], the authors approach the subject by individually considering the images of the sequence. They explore the model portability of a land-cover classifier across the sequence and weigh the efficiency of physically-based normalization techniques in mitigating the angular effects.

Overview of the Chapter

In this Chapter, we isolate and study the impact of the acquisition angle on remotely sensed images collected in an urban environment. For this purpose, we analyze two VHR sequences of multispectral images acquired within a time frame of few minutes each by the WorldView-2 satellite along an in-track collection path. The unique characteristic of these datasets is that the images represent the same urban scene under stable atmospheric, phenologic and illumination conditions: only the observation angle is varying. Our analysis is statistical: we first highlight and determine the nature of the shift in the probability distribution of the pixels caused by the increase of the off-nadir angle. For this purpose, we resort to the measure of distance between data distributions we presented in Section 4.3.2: the MMD. The analysis compares the observed distortions with respect to the type of data used, whether raw DN or atmospherically compensated data, and with reference to the application of traditional HM strategies. The statistical behavior of the spectral bands and land-cover classes is then linked to the physical properties explaining the observed angular phenomena. Subsequently, we evaluate the portability of image classification models by training a classifier on a given source image and then applying this model to all the rest of the acquisitions in the sequence (in turn considered as target images). The loss in accuracy in land-cover discrimination is analyzed with respect to the previously highlighted distortions induced by the changing view angle. We test two supervised classifiers of different nature and complexity: the LDA and the non-linear Gaussian SVM. The factors influencing the portability through the collection of images are then investigated. Here as well, we assess the ability of the statistical technique of HM in providing angle-invariant data spaces and compare it to atmospheric compensation. Thus, we evaluate the contribution of a simple yet effective relative normalization procedure with respect to a standard absolute normalization strategy.

8.2 DATA: WORLDVIEW-2 SEQUENCES OF ATLANTA AND RIO DE JANEIRO

Two in-track sequences

For the analyses presented in this Chapter, we utilized two multi-angular in-track sequences acquired by WorldView-2 over the cities of Atlanta (USA) and Rio de Janeiro (Brazil). The former, a sequence of 13 images, is described in Appendix B.4 (see page 158) while the latter, a sequence of 20 images, is presented in Appendix B.5 (see page 161). We considered multispectral imagery consisting of 8 VNIR bands (coastal, blue, green, yellow, red, red edge, NIR, NIR2). The agile satellite on-board system allows rapid imaging enabling the acquisition of a sequence of images with different view angles over the same area during a single overpass.

Angular shift

The observed dataset shift along these series of angular images is entirely due to the changing geometry of the acquisition. Indeed, since the collection of the images is quasi-simultaneous (time frame of a few minutes), factors such as changing atmospheric conditions, differences in illumination (e. g. due to the sun elevation) and seasonal effects on the vegetative cover do not impact the data acquisition. A discussion of the main reasons behind the observed angular dataset shifts, and how these effects manifest themselves in the two sequences can be found in the respective dataset description texts in the Appendix.

Preprocessing

Both image sequences were obtained in the original raw DN format, with an 11-bit dynamic range. For the present study we also considered atmospherically compensated data. The conversion to surface reflectance values has been performed using *DG-AComp*, a DigitalGlobe proprietary software allowing an automatic atmospheric compensation yielding very similar results to FLAASH [Pacifici, 2013]. Moreover, we created two additional sequences to be compared consisting of the same sets of raw DN and surface reflectance images but, in this case, after the application of the univariate HM procedure. We carried out the matching for each image in the sequence taking the CDF of the source image (specified in each of the experiments below) as the reference distribution to be reproduced.

8.3 QUANTIFICATION OF THE ANGULAR EFFECTS

Isolating the angular effects

In this Section, we analyze the spectral response of the land-cover classes with respect to the acquisition angle. As mentioned, the presented datasets provide a unique opportunity to isolate the angular effects affecting the images along an in-track multi-angle acquisition. In the following, we present the results on the detection of this angular shift by focusing on the Atlanta sequence only, whereas in Section 8.4 both sequences will be used to evaluate the model portability.

8.3.1 *A visual assessment*

Data translation due to Rayleigh scattering

Figure 8.1 provides an example of the distortions encountered in the sequence by means of a series of scatterplots relating the blue and NIR2 bands (WorldView-2 channels #2 and #8, respectively). For the raw DN data (1st column), the main apparent difference between the most nadir image (−8.5° off-nadir, image #7) and one of the two most off-nadir counterparts (+31.5° off-nadir, image #13) relates to the general translation of the data cloud observable especially along the blue axis due to Rayleigh scattering. Such a translation disappears when looking at histogram matched (2nd column) or atmospherically compensated data (3rd column). For instance, in the latter, we remark the class "shadow" appropriately exhibiting near zero reflectance values in both angular images.

Backward scattering hotspot

Nonetheless, in each data space (raw DN and after atmospheric compensation), an overall expansion of the point cloud toward brighter values (larger DNs or surface reflectances) is observed with the increase of the off-nadir acquisition angle. This effect can be attributed to the geometry of the acquisition (see Fig. B.5 on page 160), whereby the sensor is imaging a scene directly illuminated by the sun, with elements like trees being less affected by self-shadowing (backward scattering hotspot). The relative normalization of the overall distributions (2nd and 4th column) helps in mitigating this shift, as it is particularly visible for the class "grass".

8.3.2 *Experimental setup*

Source and target datasets, MMD parameters

To quantitatively assess the degree of distortion along the sequence, we measured the distance between the probability distributions of the images of the sequence with the MMD. This metric has been computed between pixels extracted from the most nadir image (−8.5° off-nadir angle), considered always as the source image (represented with a dataset X_S of size n_S), and each image of the sequence (including the source image itself), taken in turn as target images (represented with datasets X_T of size n_T). We sampled the sequence to obtain 10 independent realizations of sets with 100 pixels per class from each image. The MMD has then been computed as a separate measure for each individual class (X_S and X_T of size $n_S = n_T = 100$ pixels belonging to the same class) and as an overall measure describing

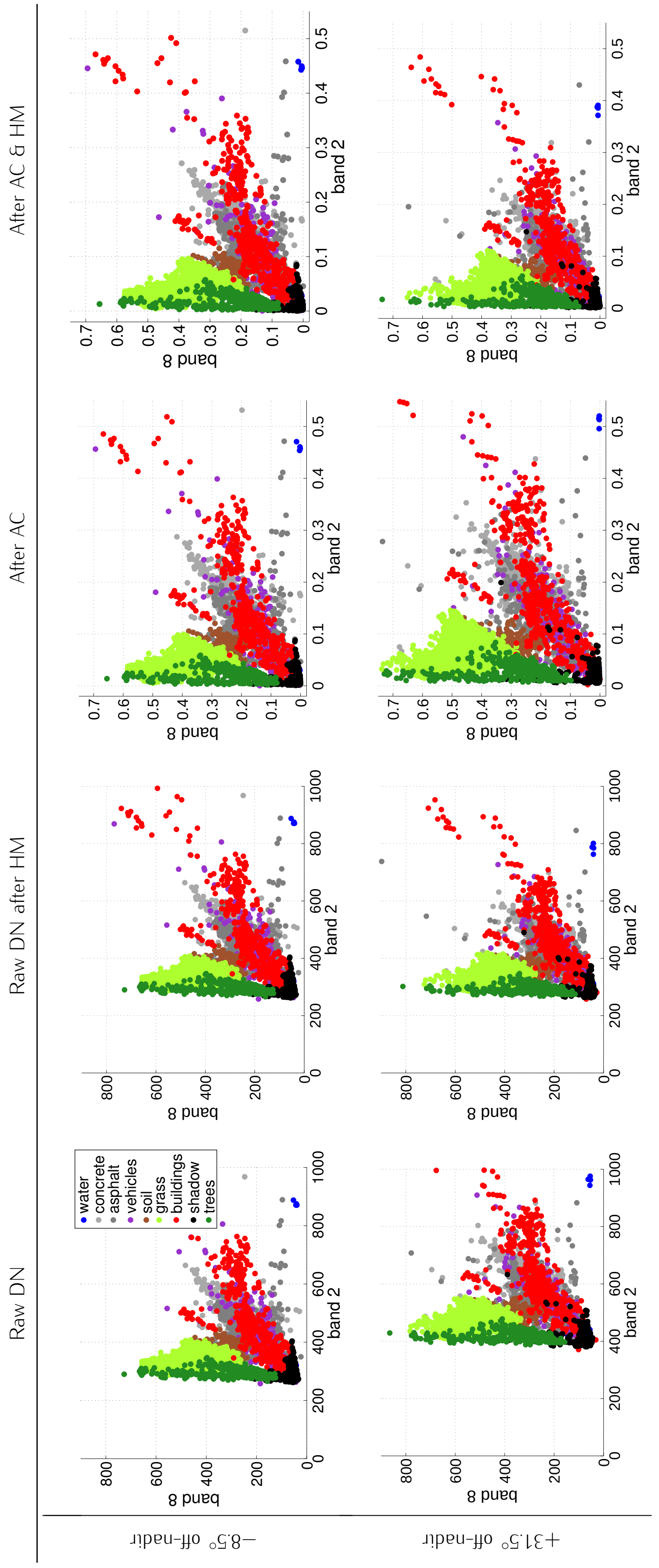


Figure 8.1: Atlanta dataset: scatterplots in the blue (band 2) vs. NIR2 (band 8) space of the image at (top) −8.5° off-nadir angle and (bottom) +31.5° off-nadir angle. (1st column) Raw DN data. (2nd column) Raw DN data after HM taking the image at −8.5° as reference. (3rd column) Surface reflectances obtained after atmospheric compensation (AC). (4th column) Surface reflectances obtained after AC and after HM taking the image at −8.5° as reference.

the distortion of the complete set of classes ("all classes" case with X_S and X_T of size $n_S = n_T = 100 \cdot 9 = 900$ pixels from the 9 classes). Additionally, besides assessing the angular evolution by considering all the bands together (X_S and X_T of dimension $d = 8$, as the number of WorldView-2 channels), we also gauged the shift for each single spectral band by calculating univariate MMD scores. A Gaussian RBF kernel with a σ width parameter equal to the square root of the number of variables involved (1 or 8) has been chosen to obtain the kernel matrices of Eq. (4.7) used by MMD (see Section 4.3.2).

Statistical normalization

For these analyses, both the raw DN and the atmospherically compensated sequences have been statistically normalized by dividing each pixel value of all the images in the sequence by the maximal value observed among the labeled pixels of the source image over the 8 bands. The purpose is to ensure a fair comparison of the trends across the data spaces while keeping the shape of the spectral signatures unaltered (the relationship between the band magnitudes is not changed).

8.3.3 *Results and discussion*

Figure 8.2 illustrates the MMD plots obtained. We present the results in a logarithmic scale for the four sequences described above: 1) raw DN data, 2) raw DN data after HM, 3) surface reflectances obtained after atmospheric compensation, and 4) atmospherically compensated data after HM. Moreover, additional to the "all classes" case, we report separate graphs for certain classes of interest, namely "shadow", "trees" and "asphalt".

We begin by first examining the plots considering all the classes together (1st row). When working with raw DN data (1st column), the statistical distance between distributions increases as the off-nadir angle increases. In agreement with the observations in Longbotham et al. [2012a], short wavelength bands subject to Rayleigh scattering (coastal, blue and green) display stronger shift trends due to the increasing up-scattered path radiance at lower satellite elevations (high off-nadir angles). The other striking tendency relates to the larger MMD values in the solar backward scatter region (off-nadir angle > 9.5°) than in the forward solar scattering region (off-nadir angle $\leq$ 9.5°). Such a phenomenon can be related to the decreased self-shadowing effects for the objects in the scene when imaging in the backward scattering region. In the raw DN histogram matched sequence (2nd column) and in the atmospherically compensated sequence (3rd column), a clear general decrease in the MMD values (by at least a factor of 10) is noticed and, above all, a decisive reduction of the shift for the bands of shortest wavelength can be identified. In particular, we draw the attention to the almost flat curve for the coastal band in the plot for histogram matched raw DNs, which is coherent with the successful correction of the data cloud translation noticed in Fig. 8.1. Applying HM to the already atmospherically compensated sequence (4th column) allows for an even stronger reduction of

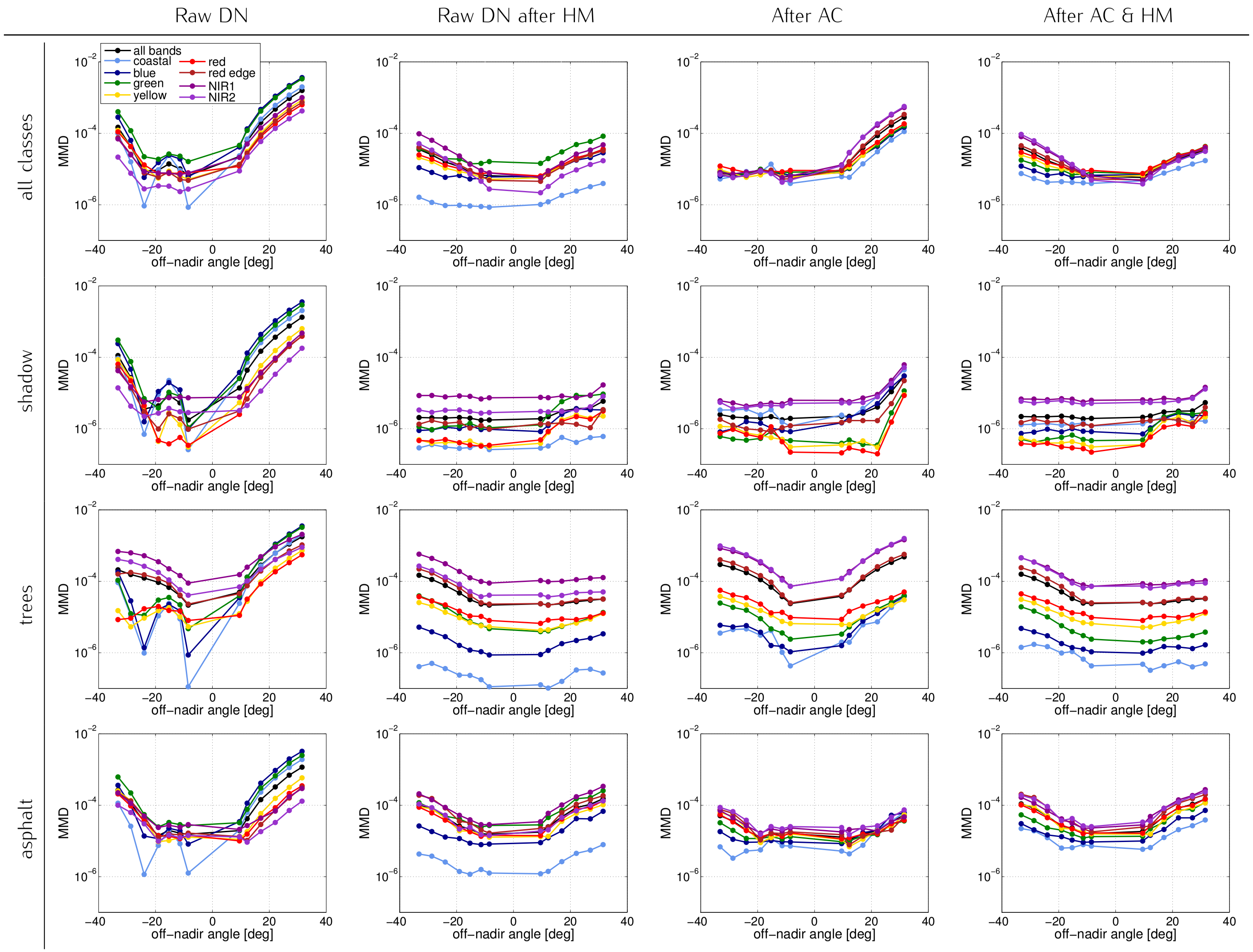


Figure 8.2: Atlanta dataset: MMD plots in logarithmic scale, by band and for all the bands together, between labeled pixels of the image at −8.5° off-nadir angle (source image) and each angular image (average of 10 experiments). Comparison of the sequences of (1st column) raw DN data, (2nd column) raw DN data after HM taking the source image as reference, (3rd column) surface reflectances obtained after atmospheric compensation (AC) and (4th column) surface reflectances obtained after AC and after HM taking the source image as reference. Results separately plotted for (1st row) "all classes", (2nd row) class "shadow", (3rd row) class "trees" and (4th row) class "asphalt".

the general shift in the backward scattering region while slightly inflating it in the forward scattering region.

Taking a closer look at specific classes, for the "shadow" class (2nd row) the reduction of the shift is appropriately obtained both through physically- and statistically-based corrections. Indeed, the consequences of the scattering of short wavelength radiation that affect the raw DN MMD plot have successfully been mitigated both with an absolute atmospheric compensation and with a relative HM. For completeness, we draw the attention to the much smaller residual off-nadir shift (close to one order of magnitude) in the atmospheric compensation plots for the "shadow" class with respect to the "all classes" case, attesting thus that the distortion for this class was essentially due to the up-scattered path radiance.

When analyzing the class "trees" (3rd row), we remark sensibly higher overall MMD values as well as the persistence of a strong angular divergence even after the corrections. This is particularly visible for the WorldView-2 channels designed to highlight vegetation properties (NIR, NIR2, red edge). This matches well with the observation that, once the atmospheric effects are removed, the response of vegetation in the longer wavelength bands is the most affected by the off-nadir acquisitions because of the illumination hotspot effects. Again, we notice the ability of HM (2nd and 4th column) in handling these distortions by suitably normalizing the data spaces in the backward scattering region of the sequence. Indeed, in correspondence of positive off-nadir angles, we notice a stark reduction of the angular shift for the NIR and NIR2 bands. This is the outcome of the compensation of the illumination/self-shadowing effects observed in Section 8.3.1.

For the class "asphalt" (4th row), a subtle behavior can be detected when successively applying the different normalization strategies. For this particular material, after a general decrease in the shift when converting raw DN data to surface reflectances (flatter MMD curves in the 3rd column with respect to the 1st), a HM on the atmospherically compensated sequence raises the MMD at large off-nadir angles (see 4th column plot). Such a phenomenon can be linked to the well-known harmful effect that HM can have on some of the classes since the procedure acts on the global CDF instead of on the CDFs of each class.

8.4 CLASSIFICATION MODEL PORTABILITY ASSESSMENT

Objectives

In this Section, we study the portability of classification models built on a single acquisition (source image), when used to predict the land-cover on all the images of the sequence (target images). With this exercise, we are interested in studying the consequences of the previously highlighted angular dataset shift on the accuracy of thematic classification. We analyze the compensation brought by statistical and physical normalization strategies, as well as the possible benefits of their joint use. In our case study, we analyzed the model portability on both the Atlanta and Rio de Janeiro sequences.

Table 8.1: The four factors analyzed during the model portability experiments with associated acronyms of the options used in figure legends or captions.

Factor	Cases
Source image	negative off-nadir angle vs. near-nadir angle vs. positive off-nadir angle
Initial data space	raw DN data ("Raw DN") vs. atmospherically compensated data ("AC")
Histogram Matching	matching not applied ("No HM") vs. matching applied ("With HM")
Type of classifier	Linear Discriminant Analysis ("LDA") vs. SVM with Gaussian kernel ("Gaussian SVM")

8.4.1 *Experimental setup*

Portability from three source images

For both datasets, we carried out 3 separate experiments, each one considering a specific image as the source domain where classifiers are trained. For the Atlanta sequence, the source images have been chosen as the acquisitions at $-24°$ (image #3), $-8.5°$ (image #7, the most nadir) and $+17°$ (image #10). For the Rio de Janeiro dataset, the image considered as the source were those at $-38.8°$ (image #5), $-6.1°$ (image #9, the most nadir) and $+39.5°$ (image #16). Such a setup has been adopted to investigate the possible asymmetry in the ability to transfer a classifier trained in angular regions subject to distinct solar scattering patterns (forward and backward scattering). As mentioned, the target image on which to test the model was drawn sequentially from the complete set of acquisitions to obtain a sequence of 13 (for Atlanta) or 20 (for Rio de Janeiro) accuracy values. We considered the estimated Kappa statistic as overall precision metric and F-measure for class-specific accuracies (see Appendix A). We made use of 10 independent training sets of 100 randomly selected pixels per class, whereas the test set was composed of all the remaining labeled pixels ($84,055$ for Atlanta and $17,963$ on average for Rio de Janeiro).

Influence of three factors

Besides the location of the source image, we examined the influence of three additional factors whose respective cases are detailed in Table 8.1: 1) the initial data space, 2) the application of HM, 3) the type of classifier used. For both initial data spaces, the dataset normalization and the application of HM was executed following the same procedures described in Section 8.3.2 and in Section 8.2.

Two classifiers

With the purpose of underlining the generalization abilities of two fundamentally different classifiers, we decided to make use of LDA and of the more sophisticated Gaussian SVM. The parameters of the SVM have been tuned by 5-fold cross-validation with a comprehensive grid search. For the penalty parameter C, the search has been carried out in the range $\{10^{-1}, \ldots, 10^{4}\}$ while for the Gaussian kernel width parameter σ, we searched the space $\{0.2, \ldots, 5\}$ times the median distance among training data points.

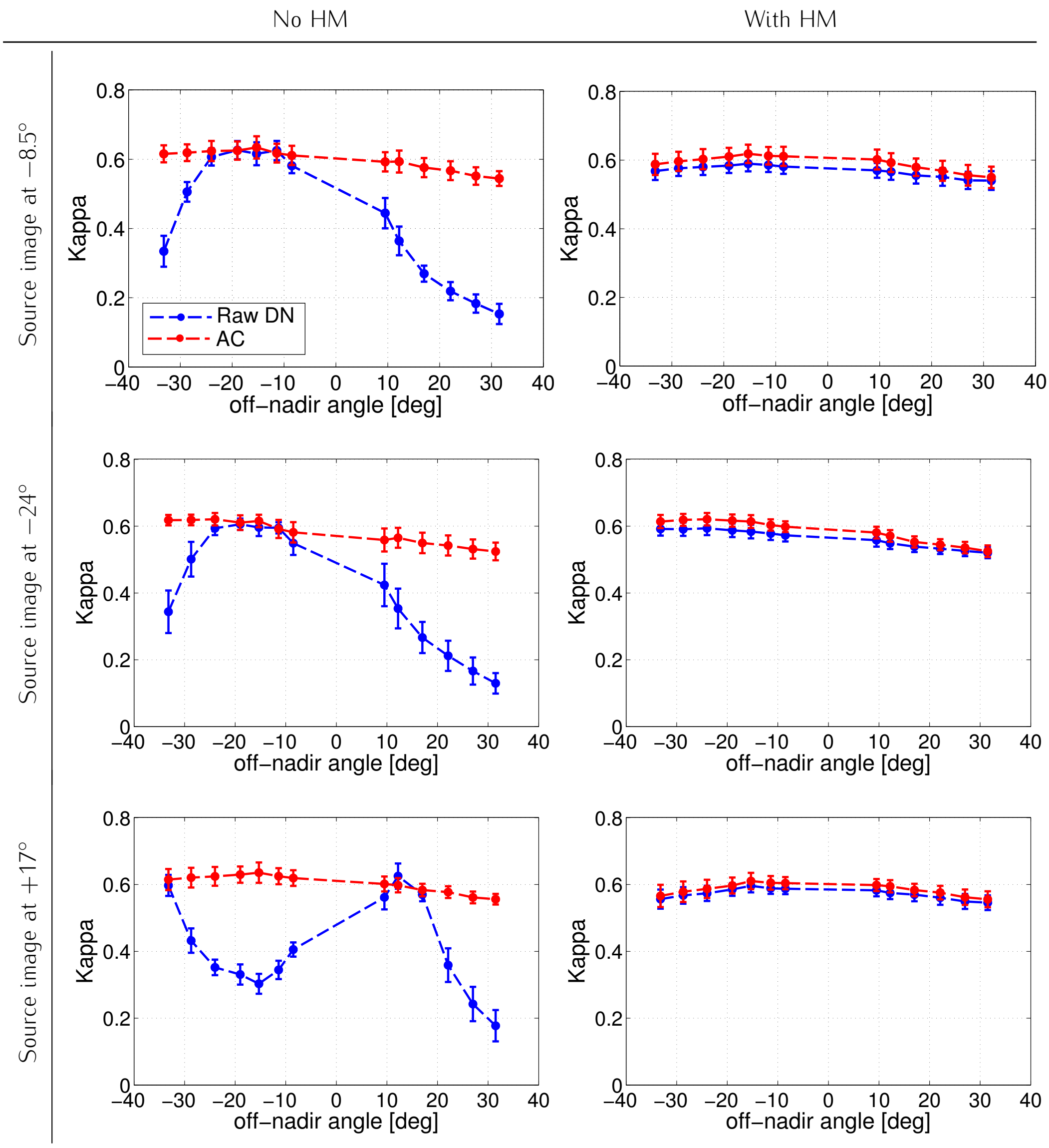


Figure 8.3: Atlanta dataset: assessment (average Kappa statistic with standard deviation over 10 experiments) of LDA model portability to all the images in the sequence in turn (target images) from the source image at (top) −8.5° off-nadir, (center) −24° off-nadir and (bottom) +17° off-nadir. Results are presented separately for (left) original unmatched sequences and (right) histogram matched sequences. Data space: "Raw DN" (raw DN data) vs. "AC" (atmospherically compensated data).

8.4.2 *Results and discussion*

8.4.2.1 *Atlanta dataset*

We first analyze the results obtained with the LDA classifier when trained in different regions of the angular sequence. Figure 8.3 illustrates the performances (Kappa statistic) of the classifiers in the various scenarios described above. Without HM (left column), portability appears to be very much dependent on the application of an appropriate atmospheric compensation (red curve). In fact, with raw data (blue curve), when training on the near-nadir image at −8.5° off-nadir or on that at −24° off-nadir, the portability suffers acutely when the off-nadir angle increases, especially in the backward scattering region (angles > 10°). This last observation properly correlates with the larger shift detected for this region of the angular collection on the MMD plots (top left plot of Fig. 8.2). Moreover, we remark very similar behaviors for these two source images, a trend explained by the fact that both acquisitions lie in the solar forward scattering region, where MMD values were modest. On the other hand, when training on the image at +17° off-nadir, the obtained plot is completely different. Besides noticing the expected Kappa peak in the vicinity of +17°, we detect a sheer drop going toward the forward scattering region. In this direction, the lowest point is attained for image #5 (−15.3°), which corresponds to the acquisition with the most pronounced forward scattering effect, i. e. a collection performed directly opposite the sun with respect to the imaged area (specular reflection).

Using images bearing matched histograms (right column), when considering the raw DN case, the classifier reaches satisfactory performances competing with those observed with atmospherically compensated data. The portability improves further if we match the histograms of the surface reflectance data, in some cases (source image at −24° off-nadir) even exceeding the unmatched atmospherically compensated profile (red curve in left column plots). In this respect, we point out that HM is favored by the specificities of this case study: the scene remains unchanged, without any change in the proportion of the land-covers or without the appearance of any new class. Those events would sensibly modify the shape of the global probability distribution, lessening therefore the appropriateness of the matching at the class level. In a context involving geographically disjoint scenes, the absolute atmospheric compensation approach should show a heavier gain in model portability.

In general, after the physical correction or the statistical matching of the images, we notice a minimal loss in accuracy when moving from the respective source images to the most off-nadir target images in the collection.

Figure 8.4 illustrates the portability performances obtained with a Gaussian SVM trained on the most nadir source image (−8.5° off-nadir). The main difference with the corresponding LDA plots lies in the higher overall performances for the SVM, more than 0.2 Kappa points superior to that of the parametric linear model across large parts of the angular domain. Considering the unmatched sequences (left), a clear dependence on the compensation of the atmospheric effects appears as the model moves off-nadir. As noticed above for LDA, the change in illumination/shadowing conditions towards the angles > 10° (backward scattering regime) hampers massively the classifier. With histogram matched sequences (right), the general performance of the SVM in this critical region of the sequence improves even more (along with

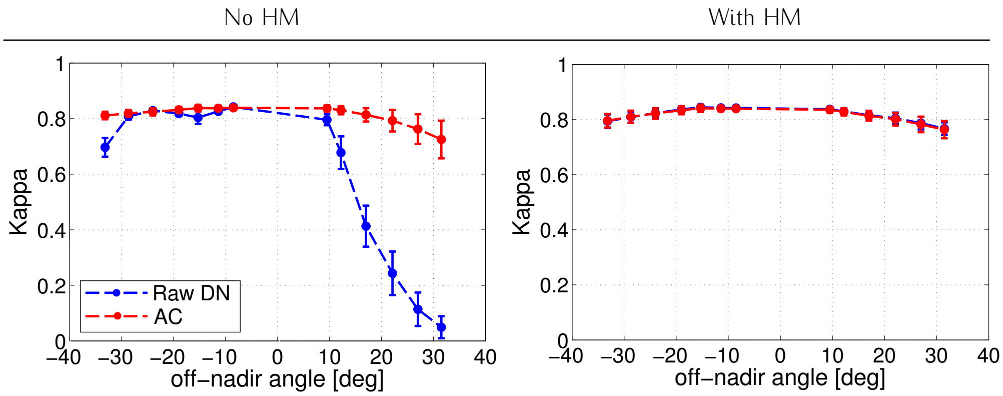


Figure 8.4: Atlanta dataset: assessment (average Kappa statistic with standard deviation over 10 experiments) of Gaussian SVM model portability from the image at −8.5° off-nadir (source image) to all the images in the sequence in turn (target images). Results are presented separately for (left) original unmatched sequences and (right) histogram matched sequences. Data space: "Raw DN" (raw DN data) vs. "AC" (atmospherically compensated data).

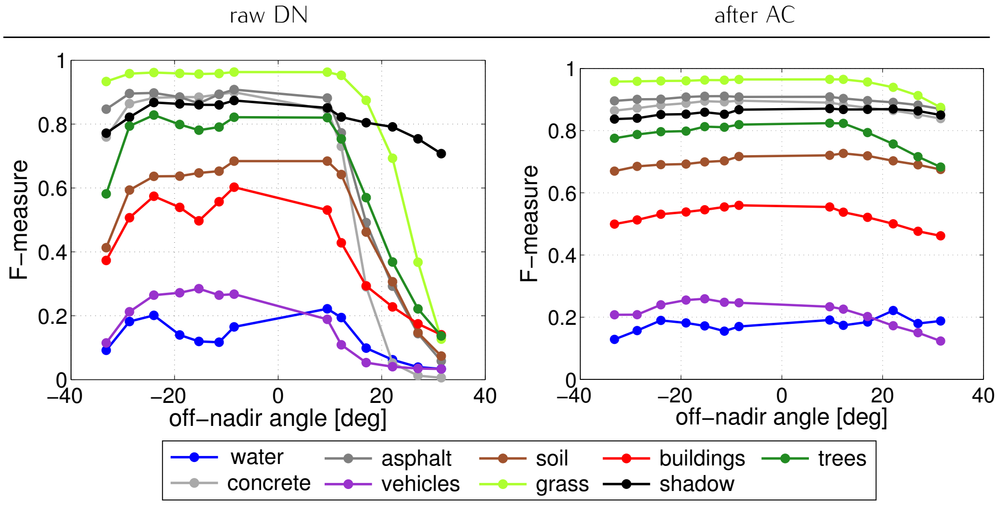


Figure 8.5: Atlanta dataset: evolution of the F-measure (average of 10 experiments) for each class in the experiment with the image at −8.5° off-nadir as source image. We considered (left) raw DN and (right) atmospherically compensated sequences (AC). A Gaussian SVM classifier and a setting without any preceding HM have been chosen.

a reduction in the variability). However, in this case, throughout the angular domain, no distinction can be made between the behavior with raw DN or with atmospherically compensated data.

Figure 8.5 allows us to investigate in more details the model portability performances by having a look at class-specific trends. We focus on the experiment without HM where a Gaussian SVM classifier was trained on the image at −8.5° off-nadir (corresponding to the left panel of Fig. 8.4). From the raw DN plot (left plot of Fig. 8.5), we understand how all the land-covers suffer the increase in off-nadir angle in the backward scattering region. The only class not being heavily affected is the class "shadows". Working with atmospherically compensated data (right plot of Fig. 8.5), a clear benefit for all the thematic classes is observed. The series of F-measure figures across the entire angular domain are now more stable, with minimal accuracy losses when moving from the near-nadir acquisition to the off-nadir counterparts. Moreover, we point out how, in both plots, very similar types of materials such as "concrete" and "asphalt" are appropriately discriminated by the SVM. On the other hand, we remark the apparent difficulty of the model in correctly detecting the classes "water" and "vehicles", those being associated with small objects (pools and parked cars) with highly variable spectral signatures.

8.4.2.2 *Rio de Janeiro dataset*

The Kappa statistic plots obtained by training the LDA model in three different angular locations of the Rio de Janeiro sequence are presented in Fig. 8.6. Considering the case without HM (left column), the main observation is that, for this sequence as well, atmospheric compensation (red curve) is crucial to achieve a good portability.

Examining the overall shape of the plots, the other main noticeable trend is that, the evolution of the Kappa curves associated with unmatched raw DN data (blue curves) presents a striking difference if compared to the Atlanta sequence (left column of Fig. 8.3). As a matter of fact, a clear symmetry with respect to the nadir is visible in the present case. When setting the acquisition with the lowest absolute off-nadir angle (image at $-6.1°$) as the source image, we notice an almost equivalent decay in accuracy on each side of the angular sequence for the raw DN data. The motivation for such a distinct behavior can be traced back to the acquisition geometry. Indeed, during this collection overpass, the sun was perpendicular to the satellite flight path and could illuminate the scene with similar shadowing effects on each side of the sequence.

The center and bottom rows of Fig. 8.6 report the results obtained when the images with highly slanted geometries ($-38.8°$ and $+39.5°$ off-nadir) have been used as the source domain. In both situations, the general shape of the Kappa statistic curves reveals moderate accuracies in the central region of the acquisition (off-nadir angle between $-30°$ and $+30°$), then increasing when moving toward both the $-40°$ and $+40°$ angles (either the angular region close to where the classifier has been trained or its symmetrical opposite). This matches the considerations about the similar illumination/shadowing conditions existing in these off-nadir regions of the Rio de Janeiro sequence, leading to an adequate portability among them.

Overall, the Kappa values are higher on this dataset due to the lower number of classes. Moreover, for absolute off-nadir angles $> 40°/45°$, the decrease in accuracy is associated with an extremely large variability among the experiments. This behavior was not observable in the Atlanta sequence, since only the Rio de Janeiro dataset features such oblique look angles.

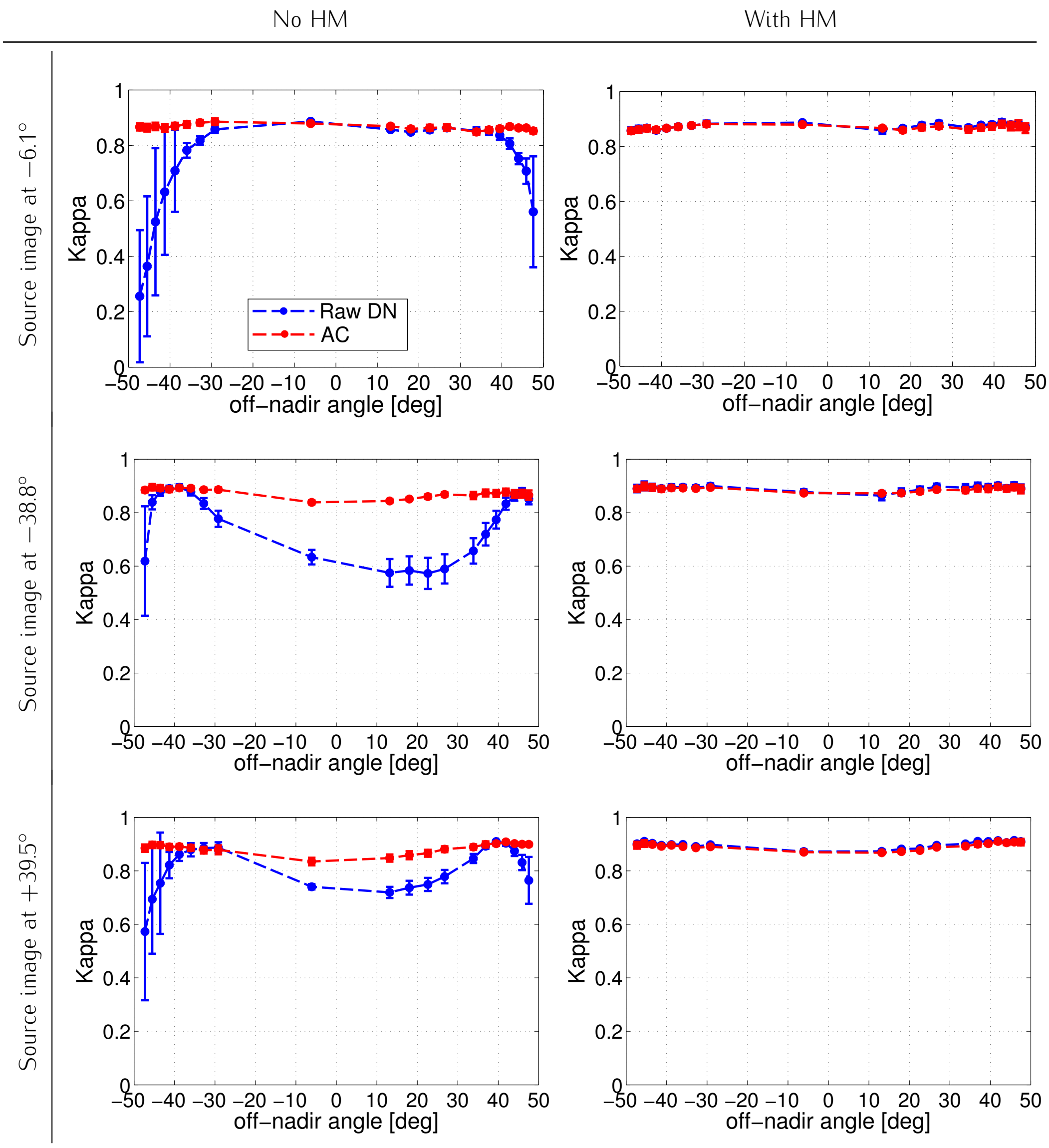


Figure 8.6: Rio de Janeiro dataset: assessment (average Kappa statistic with standard deviation over 10 experiments) of LDA model portability to all the images in the sequence in turn (target images) from the source image at (top) −6.1° off-nadir, (center) −38.8° off-nadir and (bottom) +39.5° off-nadir. Results are presented separately for (left) original unmatched sequences and (right) histogram matched sequences. Data space: "Raw DN" (raw DN data) vs. "AC" (atmospherically compensated data).

Results with HM (right column) reveal a very good portability of the LDA across the entire angular domain (almost no loss in classification accuracy), no matter the data space. In the central region of the plots for off-nadir angles of −38.8° and +39.5°, a slightly more satisfactory performance is noticed with this relative normalization technique with respect to the unmatched atmospherically compensated sequences.

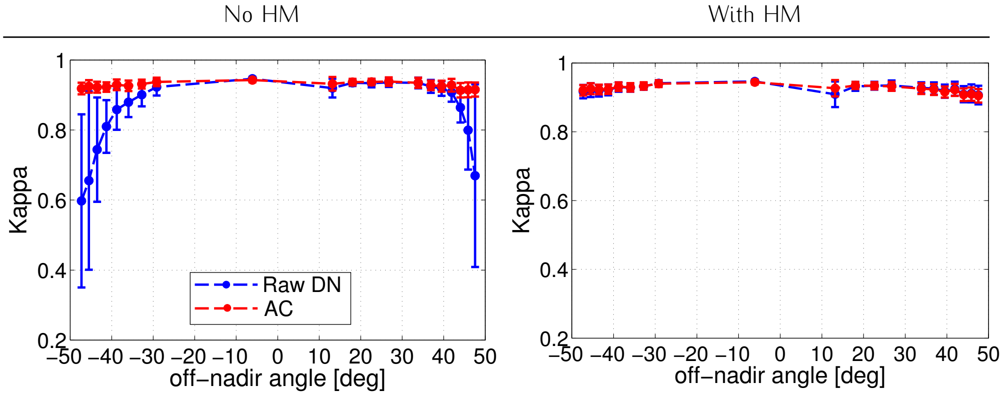


Figure 8.7: Rio de Janeiro dataset: assessment (average Kappa statistic with standard deviation over 10 experiments) of Gaussian SVM model portability from the image at −6.1° off-nadir (source image) to all the images in the sequence in turn (target images). Results are presented separately for (left) original unmatched sequences and (right) histogram matched sequences. Data space: "Raw DN" (raw DN data) vs. "AC" (atmospherically compensated data).

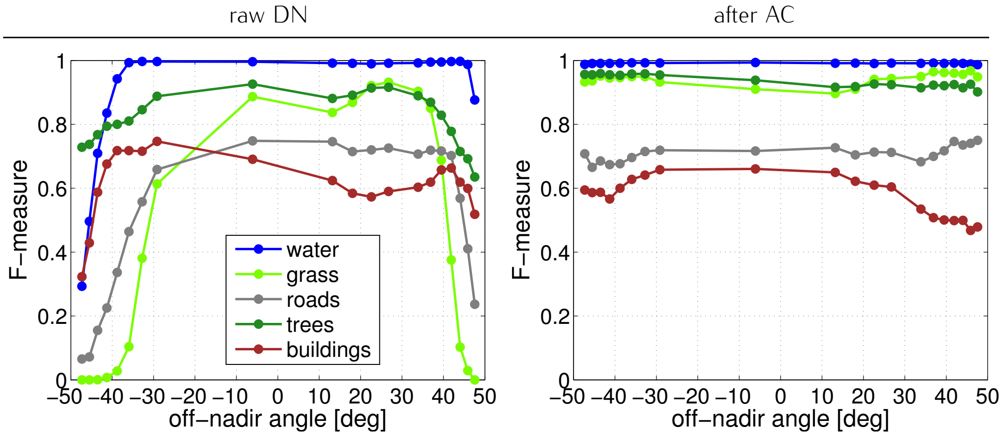


Figure 8.8: Rio de Janeiro dataset: evolution of the F-measure (average of 10 experiments) for each class in the experiment with the image at −6.1° off-nadir as source image. We considered (left) raw DN and (right) atmospherically compensated sequences (AC). A LDA classifier and a setting without any preceding HM have been chosen.

For the analysis of the results obtained with the Gaussian SVM classifier (Fig. 8.7), we only present the experiments with the near-nadir source image (−6.1°). As expected, the most accurate thematic maps are produced when using the classifier on the atmospherically compensated (red curve of the left plot) or on the histogram matched sequences (both curves of the right plot). These Gaussian SVMs show an average precision stable at Kappa > 0.9 throughout the sequence.

With Fig. 8.8, we break down the portability results by land-cover type in order to highlight the benefits of the transformation to surface reflectance values. To this end, we retained the experiment using the LDA classifier trained on the image at −6.1° off-nadir in the setting without HM (corresponding to the top-left panel of Fig. 8.6). On the raw DN plot (left plot of Fig. 8.8), the evolution of the F-measure points out the classes "water", "grass" and "roads" as those heavily suffering the skewed angular acquisitions. Instead, land-covers such as "trees" seem to be less affected. As soon as we turn to the physically normalized space (right plot of Fig. 8.8), the improvement is notable for all the classes. The ability to compensate the distortions caused by atmospheric effects becomes apparent at large off-nadir angles, in particular for the critical classes cited above.

8.5 CONCLUSIONS

Main achievements

In this study, two in-track VHR multispectral sequences were used to find means to evaluate and possibly correct the shifts in data distributions caused differences in acquisition angle. These quasi-simultaneous collections of multiple images allowed us to isolate the effects caused by the acquisition geometry. The distortions induced by physical phenomena controlling the radiation transfer could be quantified thanks to a robust statistical measure of distance between probability distributions, the MMD. By means of this insightful non-parametric statistic, we could highlight key effects such as the increasing Rayleigh scattering when imaging at high off-nadir angles and its disappearance when working with surface reflectance data. By testing the model portability of classifiers across the sequence, we could describe the evolution of the thematic classification accuracy through the angular domain with considerations about the influence of the location of the source image in the angular sequence. The experimental trends agree with the observations related to the dataset shift highlighted in the first place.

Atmospheric compensation & HM: suitable normalizations

Additionally, we studied the influence of classic preprocessing techniques on the generalization abilities of the models. The basic trends we remarked can be summarized as follows. On the one hand, a precise atmospheric compensation provided images with similar radiometric characteristics over the entire angular domain. The residual shift can be imputed to BRDF or observational solar cross-section effects not accounted for with the transformation into surface reflectance values. On the other hand, good results in compensating for the angular divergence have also been observed by applying a band-by-band matching of the histograms. Such an approach, even though expected to be less effective on images coming from separate spatial locations, proved able to overcome the shortcomings of the change in acquisition angle.

Physics & machine learning: complementarity

The study has also underlined the complementarity of the physical and machine learning approaches. Indeed, after an absolute normalization by atmospheric compensation, remarkable portability performances were obtained by employing a state-of-the-art kernel-based method. In this respect, we emphasize the key point related to the non-linearity of the cross-image knowledge transfer process. The empirical results we provided revealed that, once appropriate radiometric corrections are applied, by making use of a non-linear Gaussian SVM in the classification step we obtain an adaptive land-cover classification system largely immune to the effects of the view angle. The linear and parametric approach of LDA appeared much more prone to fail in critical angular shift situations.

Angle-invariant data spaces for large-scale mapping

In general, appropriately chosen normalization approaches ensuring angle-invariant data spaces, combined with the most flexible and portable models, allow to extend the classification rules over multiple images acquired with different geometries. This could ultimately enable a successful large-scale land-cover mapping. As a conclusive remark, we demonstrated that, taking the best of both worlds, the joint use of physically-based atmospheric compensation approaches along with statistical/machine learning matching and classification techniques allows to attain the desired model portability in multi-angle VHR sequences.

What's next

Future research directions will be focused on the analysis of the generalization abilities of land-cover models when working with composite multi-angle data, i. e. an ensemble of image acquisitions carried out during multiple satellite overpasses. Such images will thus be characterized by markedly different satellite and sun positions (elevations and azimuths), since they were not exclusively collected over the same in-track acquisition path. The dataset shift due to the geometry of the collections could then be studied over the entire azimuth-elevation space and not only on a cross-section of it.

9 CROSS-IMAGE SYNTHESIS WITH DICTIONARIES

Outline: *This Chapter studies an approach based on Dictionary Learning which enables the alignment of the sparse representations of two images. A linear transformation is derived thanks to an algorithm simultaneously learning the image-specific dictionaries and the mapping function bridging them via their respective sparse codes. In the following, Section 9.1 discusses the advantages of a direct cross-domain conversion of the data spaces based on sparse representations, a methodology that will then be summarized in Section 9.2. Next, in Section 9.3 we will present the particular dataset used to test the technique and the associated setting of the experiments. Section 9.4 reports the results we obtained while Section 9.5 concludes the Chapter by addressing strengths and limitations of the proposed approach.*

9.1 INTRODUCTION

Recap

The previous two Chapters revealed two noteworthy trends regarding relative normalization strategies. As we pointed out in Chapters 7 and 8, in order to lessen the dataset shift affecting the image distributions when dealing with multiple images collected under different conditions, resorting to elementary techniques such as HM can be very effective. Nonetheless, in some situations this is too simplistic [Yang and Mueller, 2007] and does not allow handling images with different number of bands (data spaces of different dimension). At the same time, as observed in Chapter 7, the capacity to project the images to an appropriate joint sub-space proved extremely beneficial to suitably match the distributions.

Direct cross-domain conversion

Another possible tempting approach consists in directly seeking a transformation able to convert the data space of one image to that of another one. The absence of an intermediary sub-space ensures that only one of the two images has to be transformed. Moreover, by devising a method flexible enough, it would be very useful if images with different numbers of spectral bands could be treated by the procedure. This would eventually enable cross-sensor transformations.

This Chapter will appear in:

> G. Matasci, F. de Morsier, M. Kanevski, and D. Tuia. Domain adaptation in remote sensing through cross-image synthesis with dictionaries. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Québec City, Canada, 2014.

Suitability of DL

To this end, in this Chapter, we take advantage of the DL framework presented in Section 3.6. As reviewed, DL is a rising field of investigation in hyperspectral remote sensing, where it has shown promising results in classification with compact models relying on the hypothesis of sparsity. Likewise, the same sparsity hypothesis has proven successful in applications aiming at fusing high spatial and spectral resolutions. Song et al. [2014] suggest learning a dictionary-pair to describe a multispectral and a hyperspectral image through dictionaries possessing the same number of atoms. Subsequently, they seek a single matrix of sparse codes reconstructing the pixel signals of the two images. This matrix allows to link the spectral properties of each material in the low spectral resolution image to those in the high spectral resolution acquisition. Such a bridge ultimately enables the synthesis of pixels bearing both a high spatial and a high spectral resolution.

Overview of the Chapter

Hereafter, we propose to align sparse representations based on dictionaries defined on the images of interest in order to perform the adaptation. The key idea of this cross-domain image synthesis is that the pixels of a remote sensing image can be converted to the data space of another related image by means of a linear transformation [Wang et al., 2012]. The algorithm simultaneously learns a dictionary pair (one per image) and a mapping function from one to the other. The dictionaries characterize the structure of the domains, while the mapping encodes the relation between them. Once the transformation is found, a cross-domain synthesis can be carried out to convert an image into another, easing thus the DA task at hand. This approach can be related to that of [Wang and Mahadevan, 2011] in the sense that the latter also enables the user to reciprocally translate the image data spaces. However, in their case, this is done through an intermediate step involving a mapping to a common latent space.

9.2 DICTIONARY LEARNING FOR CROSS-IMAGE SYNTHESIS

9.2.1 *Problem formulation*

Two generic domains

The transformation we study in this Chapter can be applied in both directions (from the source image to the target image or inversely). For this reason, the notions of source and target domain will be introduced only for the classification phase. In the following, data spaces $\mathcal{X}$ and $\mathcal{Y}$, data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, samples $\boldsymbol{x}$ and $\boldsymbol{y}$ as well as the associated "$\cdot_x$" and "$\cdot_y$" subscripts denote elements referring to two generic but distinct domains $\mathcal{D}_x$ and $\mathcal{D}_y$. We point out that, therefore, $\mathcal{Y}$ is not considered here as the output space of the class labels.

The principle

Bearing this in mind, the task of cross-domain image synthesis consists in finding an invertible mapping $f(\cdot)$ allowing to translate the data space $\mathcal{X}$ of a first image, i. e. the domain $\mathcal{D}_x$, into the data space $\mathcal{Y}$ of a second image, i. e. the domain $\mathcal{D}_y$, and inversely: $\mathcal{Y} = f(\mathcal{X}),\ \mathcal{X} = f^{-1}(\mathcal{Y})$.

Inputs of the algorithm

The algorithm requires paired training data matrices $\boldsymbol{X} \in \mathbb{R}^{d_x \times n}$ and $\boldsymbol{Y} \in \mathbb{R}^{d_y \times n}$ composed of n signals belonging to the first and second domain, respectively[1]. Note that the dimension of the two data spaces can differ, i.e. $d_x \neq d_y$. The mapping function is determined by seeking a conversion matrix $\boldsymbol{W}$ aligning the sparse coding coefficients $\boldsymbol{C}_x \in \mathbb{R}^{K \times n}$ over dictionary $\boldsymbol{D}_x \in \mathbb{R}^{d_x \times K}$ to those in the other domain, i.e. $\boldsymbol{C}_y \in \mathbb{R}^{K \times n}$ over dictionary $\boldsymbol{D}_y \in \mathbb{R}^{d_y \times K}$ [Wang et al., 2012].

Optimization problem

The optimization problem to jointly retrieve the dictionaries and the mapping matrix $\boldsymbol{W}$ is the following:

$$\begin{aligned} \min_{\{\boldsymbol{D}_x,\boldsymbol{D}_y,\boldsymbol{W}\}} \quad & \Big\{ \left\|\boldsymbol{X}-\boldsymbol{D}_x\boldsymbol{C}_x\right\|_F^2 + \left\|\boldsymbol{Y}-\boldsymbol{D}_y\boldsymbol{C}_y\right\|_F^2 \\ & + \eta \left\|\boldsymbol{C}_y - \boldsymbol{W}\boldsymbol{C}_x\right\|_F^2 + \zeta \left\|\boldsymbol{W}\right\|_F^2 \Big\} \\ \text{s.t.} \quad & \left\|\boldsymbol{c}_{x,i}\right\|_0 \leq s_x, \quad \left\|\boldsymbol{c}_{y,i}\right\|_0 \leq s_y \quad , \\ & \left\|\boldsymbol{d}_{x,i}\right\|_2 \leq 1, \quad \left\|\boldsymbol{d}_{y,i}\right\|_2 \leq 1 \quad \forall i \ , \end{aligned} \tag{9.1}$$

where η is a tradeoff parameter, ζ is a regularization parameter and s_x, s_y are the sparsity levels tolerated for each dictionary. Vectors $\boldsymbol{c}_{x,i}, \boldsymbol{c}_{y,i}$ are the sparse codes constituting $\boldsymbol{C}_x, \boldsymbol{C}_y$ while vectors $\boldsymbol{d}_{x,i}, \boldsymbol{d}_{y,i}$ are atoms of $\boldsymbol{D}_x, \boldsymbol{D}_y$, respectively. Concretely, the first two terms of (9.1) represent the reconstruction error in the two domains, the third term relates to the linear mapping error between the two domains, while the constraints ensure the sparsity of the solution.

9.2.2 *Training step*

Three sub-problems

The above optimization problem is solved by splitting (9.1) into three separate sub-problems:

- the sparse coding for the training samples $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$,
- the update of the dictionaries of the two domains $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$,
- the update of the mapping matrix $\boldsymbol{W}$.

Sparse coding

The first sub-problem needs an initialization of both the mapping matrix and the dictionaries. We recall that the mapping can be carried out in both directions ($\mathcal{X} \rightarrow \mathcal{Y}$ and $\mathcal{Y} \rightarrow \mathcal{X}$). Thus, in the following joint optimization problem (9.2), we will be specifically referring to $\boldsymbol{W}$ with $\boldsymbol{W}_{x \rightarrow y}$, denoting the matrix executing the mapping of the pixels from data space $\mathcal{X}$ to data space $\mathcal{Y}$, whereas we will use $\boldsymbol{W}_{y \rightarrow x}$ to refer to the matrix carrying out the inverse task. These two matrices can be initialized as the identity matrix. The dictionaries $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$ can be initialized independently in each domain by K-SVD [Aharon et al., 2005], an algorithm also returning initial guesses for

[1] Please remark that in this Chapter, to meet the DL notation, data matrices usually consisting of n rows and d columns are transposed, i.e. of size $d \times n$.

the corresponding sparse codes $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$. Afterwards, these same sparse codes can be jointly recomputed through these two minimization problems:

$$\begin{aligned} \min_{\{\boldsymbol{C}_x\}} \quad & \left\{ \left\|\boldsymbol{X} - \boldsymbol{D}_x\boldsymbol{C}_x\right\|_F^2 + \eta \left\|\boldsymbol{C}_y - \boldsymbol{W}_{x\to y}\boldsymbol{C}_x\right\|_F^2 \right\} \\ \text{s.t.} \quad & \left\|\boldsymbol{c}_{x,i}\right\|_0 \leq s_x \quad \forall i \quad , \\ \min_{\{\boldsymbol{C}_y\}} \quad & \left\{ \left\|\boldsymbol{Y} - \boldsymbol{D}_y\boldsymbol{C}_y\right\|_F^2 + \eta \left\|\boldsymbol{C}_x - \boldsymbol{W}_{y\to x}\boldsymbol{C}_y\right\|_F^2 \right\} \\ \text{s.t.} \quad & \left\|\boldsymbol{c}_{y,i}\right\|_0 \leq s_y \quad \forall i \ . \end{aligned} \tag{9.2}$$

Dictionary update

Now, keeping the sparse codes $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$ fixed we can update the dictionary pair $\boldsymbol{D}_x$, $\boldsymbol{D}_y$ via

$$\begin{aligned} \min_{\{\boldsymbol{D}_x,\boldsymbol{D}_y\}} \quad & \left\{ \left\|\boldsymbol{X} - \boldsymbol{D}_x\boldsymbol{C}_x\right\|_F^2 + \left\|\boldsymbol{Y} - \boldsymbol{D}_y\boldsymbol{C}_y\right\|_F^2 \right\} \\ \text{s.t.} \quad & \left\|\boldsymbol{d}_{x,i}\right\|_2 \leq 1, \quad \left\|\boldsymbol{d}_{y,i}\right\|_2 \leq 1 \quad \forall i \ . \end{aligned} \tag{9.3}$$

Concretely, this step can be implemented with a one-by-one update strategy actually separating the update of $\boldsymbol{D}_x$ and $\boldsymbol{D}_y$.

Mapping matrix update

Finally, the matrix $\boldsymbol{W}$ can be updated so as to minimize the error in the conversion of the sparse codes from one domain to the other:

$$\min_{\{\boldsymbol{W}\}} \quad \left\{ \left\|\boldsymbol{C}_y - \boldsymbol{W}\boldsymbol{C}_x\right\|_F^2 + (\zeta/\eta) \left\|\boldsymbol{W}\right\|_F^2 \right\} \ . \tag{9.4}$$

The solution to this problem can be found analytically:

$$\boldsymbol{W} = \boldsymbol{C}_y\boldsymbol{C}_x^\top(\boldsymbol{C}_x\boldsymbol{C}_x^\top + (\zeta/\eta)\boldsymbol{I})^{-1} \ . \tag{9.5}$$

9.2.3 *Synthesis step*

Synthesis of a new pixel

Once appropriate dictionaries $\boldsymbol{D}_x, \boldsymbol{D}_y$ and mapping matrix $\boldsymbol{W}$ have been jointly learned, the synthesis of a new pixel $\boldsymbol{x}_i$ from $\mathcal{X}$ to $\mathcal{Y}$ demands one last optimization problem to be solved:

$$\begin{aligned} \min_{\{\boldsymbol{a}_{x,i},\boldsymbol{a}_{y,i}\}} \quad & \Big\{ \left\|\boldsymbol{x}_i - \boldsymbol{D}_x\boldsymbol{a}_{x,i}\right\|_2^2 + \left\|\boldsymbol{y}_i - \boldsymbol{D}_y\boldsymbol{a}_{y,i}\right\|_2^2 \\ & + \eta \left\|\boldsymbol{a}_{y,i} - \boldsymbol{W}\boldsymbol{a}_{x,i}\right\|_2^2 \Big\} \\ \text{s.t.} \quad & \left\|\boldsymbol{a}_{x,i}\right\|_0 \leq s_x, \quad \left\|\boldsymbol{a}_{y,i}\right\|_0 \leq s_y \quad \forall i \ . \end{aligned} \tag{9.6}$$

The solution is obtained by alternatively updating sparse coefficients $\boldsymbol{a}_{x,i}$ and $\boldsymbol{a}_{y,i}$ after having initialized $\boldsymbol{y}_i$ as $\boldsymbol{D}_y\boldsymbol{W}\boldsymbol{a}_{x,i}$, with $\boldsymbol{a}_{x,i}$ resulting from the coding of $\boldsymbol{x}_i$ on $\boldsymbol{D}_x$. The final cross-domain synthesis is then obtained by:

$$\boldsymbol{y}_i^* = \boldsymbol{D}_y\boldsymbol{a}_{y,i} \ . \tag{9.7}$$

The newly recreated pixel $\boldsymbol{y}_i^*$, while of course still belonging to the first image, is now supposed to better reflect the characteristics of the data space $\mathcal{Y}$ of the second image. Figure 9.1 provides a graphical illustration of the principle of cross-image synthesis via DL.

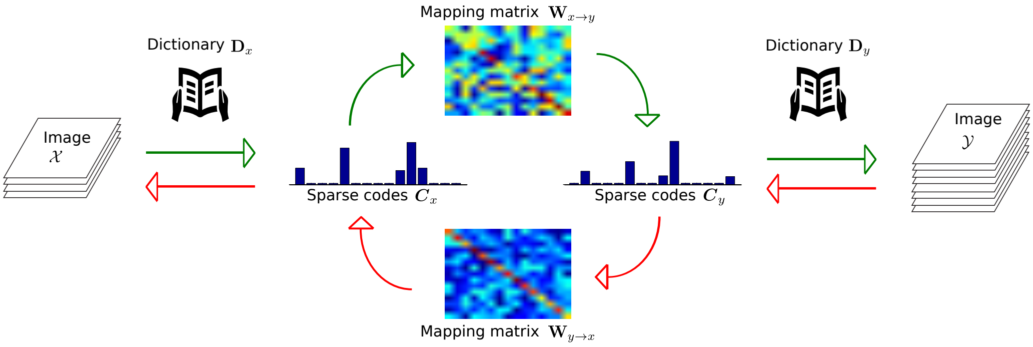

Mapping matrix $\mathbf{W}_{x\to y}$
Dictionary $\mathbf{D}_x$
Dictionary $\mathbf{D}_y$
Image $\mathcal{X}$
Image $\mathcal{Y}$
Sparse codes $\boldsymbol{C}_x$
Sparse codes $\boldsymbol{C}_y$
Mapping matrix $\mathbf{W}_{y\to x}$

Figure 9.1: Scheme of the approach for cross-image synthesis with dictionaries.

9.3 DATA AND EXPERIMENTAL SETUP

9.3.1 *WorldView-2 images of Atlanta*

Two opposite angular images

In the experiments below, we made use of two of the images belonging to the in-track multi-angle WorldView-2 acquisition over the city of Atlanta (see Appendix B.4 on page 158). We considered a first acquisition with a positive off-nadir angle of 31.5° (image #13 acquired from the south of the city) as the source image and a second image with a negative off-nadir angle of 24° (image #3 acquired from the north of the city) as the target image. For the position of the satellite during these acquisitions refer to the azimuth-elevation plots of Fig. B.5. From the initial scene visible in Fig. B.4(a) we selected a spatial subset of size 1115×1266 pixels.

Nature of the dataset shift

The dataset shift affecting these two images is associated with the distortions of the spectral signatures caused by the view angle. The source image lies in a region of the angular sequence where a strong backward scattering pattern is present (satellite on the same side of the sun with respect to the imaged area), whereas the target image lies in a region with a forward scattering regime (satellite opposite to the sun). As seen in Chapter 8, the loss in classification accuracy when porting a model across different solar scattering regions can be passably large.

Land-cover classes & preprocessing

For this study, the ground truth of the scene consisted of 45,706 pixels featuring 8 land-cover classes (see Tab B.4): "water", "concrete", "asphalt", "soil", "grass", "buildings", "shadow", "trees" (the class "vehicles" has been excluded). The acquisitions have been calibrated to surface reflectance values using the DG-AComp method (see Section 8.2). Moreover, with the goal of increasing the spatial representativeness and discriminative power of the considered signals, the initial data vector (the 8 WorldView-2 bands) has been augmented with the values of the first two principal components observed in a 5×5 neighborhood. For each component, these 24 newly added values are also sorted to guarantee invariance to rotation of the objects in the scene [Tao et al., 2014].

Table 9.1: Description of the different settings (baselines and DA strategies) used for the classification of the target image.

Name	Domain for training	Train. set [pix./class]	Detail
`TGT-large`	target	200	Uses many target pixels for training (upper bound on accuracy)
`TGT`	target	20	Uses the few available target pixels for training
`SRC`	source	200	Uses many source pixels for training (DA baseline)
`SRC-AdaptSRCtrain`	source	200	**Adaptation**: source training set converted to target domain
`SRC-AdaptTGTimage`	source	200	**Adaptation**: entire target image converted to source domain

9.3.2 *Experimental setup*

Class-specific mapping

Due to the large variability of the spectra and spatial structures encountered in the images, the definition of a global mapping valid for the entire image is highly challenging. Following an intuition similar to that of Wang et al. [2012], we decided to run the synthesis algorithm at a lower level in the hierarchy of the image: the semantic class. This means that a different $\boldsymbol{W}^{cl}$ for each class $cl \in \mathcal{C} = \{1, 2, \ldots, c\}$ is sought and that, consequently, a dedicated mapping for each land-cover is defined. To enable this option, labeled samples are needed in both images. We assume that many more labeled pixels are available in the source image, whereas just a few can be acquired in the target image (supervised DA setting).

Two mapping settings

As the projection can be applied in both directions, the pixels $\boldsymbol{x}_i$ to be synthesized by Eq. (9.7) can belong to either the source or the target domain and be projected into the other one to obtain the corresponding $\boldsymbol{y}_i^*$. For this reason, we consider two experimental settings:

- Perform a synthesis of the source training set to convert it to the target domain: this option allows the direct use of the mapping matrix $\boldsymbol{W}^{cl}$ of the respective class for each training pixel.
- Perform a synthesis of the entire target image to convert it to the source domain: this options allows to synthesize anew a complete image matching the radiometry of the source image. This option has the disadvantage of requiring the knowledge about which class-specific $\boldsymbol{W}^{cl}$ to employ for a given new target pixel to synthesize.

Compared approaches

Once both data are in the same data space, we compare the cross-image classification approaches summarized in Tab. 9.1 by assessing the performances using the ground truth of the target image. The reference accuracy (best foreseeable result) is set by the `TGT-large` method, which uses a large training set with 200 pixels per class extracted from the target image. Such a dataset is assumed to be unavailable in practical applications. Instead, `TGT` and `SRC` constitute the baseline models built from a small training set (20 pixels per class) from the target image and a large training set (200 pixels per class) from the source image, respectively. The former dataset is made up by the only ground truth assumed to be available in the target image. The latter, although being of substantial size, is not very representative of the probability distributions in the target image as it belongs to the original untransformed source image. In strategies `SRC-AdaptSRCtrain` and `SRC-AdaptTGTimage`, the cross-domain synthesis has been carried out using, for each class, training sets $\boldsymbol{X}$ and $\boldsymbol{Y}$ of size $n = 20$ pixels (all the labeled target pixels assumed to be available). For the `SRC-AdaptTGTimage` option, it is important to assign a class membership to each pixel in the target domain, in order to select the appropriate mapping matrix $\boldsymbol{W}^{cl}$ for each pixel. To do so, we used class assignments of the `TGT` strategy as initial class guesses in the target domain. We report results averaged over five random realizations of the training sets. The test set included all the pixels in the ground truth of the target image that were not used for training.

Parameters

For the proposed DL algorithm, initial dictionaries have been found by K-SVD randomly initialized with dictionary size $K = 5$. The sparsity levels s_x and s_y, which control the maximum number of atoms used for the reconstruction of a pixel, are set to 4. The regularization parameter ζ is set to 0.1 while the tradeoff parameter has been set to $\eta = 0.05$. As classifier, we used a linear SVM with a penalty parameter tuned by 5-fold cross-validation in $\{10^{-1}, \ldots, 10^{3}\}$.

9.4 RESULTS AND DISCUSSION

Figure 9.2 reports the classification performances on the target image (Kappa statistic on the test set) of the strategies described above. First, we note the very precise `TGT-large` classification, with a Kappa statistic of 0.846. The accuracy of the prediction decreases to Kappa = 0.725 if the linear SVM is trained on a set composed of 20 target samples per class only (`TGT` setting). If we try to predict the thematic classes in a cross-domain setting and without adaptation (`SRC` setting), even though the source image model relies upon 200 samples per class, the average quality of the resulting target classification maps drops to Kappa = 0.589.

Analyzing the cases were a synthesis aiming at overcoming the dataset shift is involved, we observe Kappa statistics of 0.698 and 0.711 for the `SRC-AdaptSRCtrain` and `SRC-AdaptTGTimage` approaches, respectively. These results yielded by SVM models exclusively trained on labeled source samples are quite satisfactory. Indeed, these strategies clearly improve the corresponding cross-domain approach based on the same, yet untransformed, pixels (`SRC`).

We remark that the strategies involving a cross-image synthesis are not able to outperform the `TGT` setting. On the one hand, this is due to the already known outstanding performances of the linear SVM even if trained

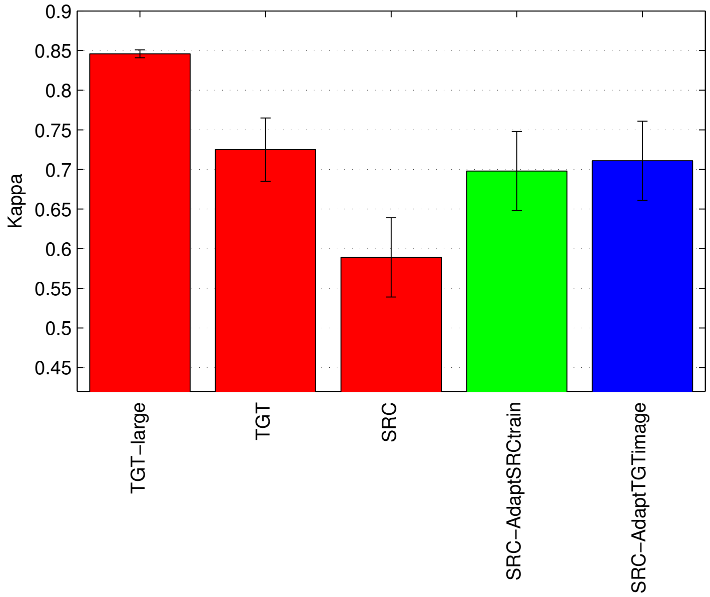


Figure 9.2: Classification performances (average and standard deviation of estimated Kappa statistic over 5 runs) obtained on the target image with the different strategies described in Tab. 9.1.

on a limited amount of labeled pixels. On the other hand, the accuracy of `SRC-AdaptTGTimage` is strongly dependent on the quality of the initial `TGT` map used to choose which class-specific mapping to use. However, we draw the attention to the fact that, in the last two cross-domain synthesis cases, the labeled target samples are not directly involved in the classification. They are only indirectly contributing to the DA task by helping the DL algorithm in designing the transform.

9.5 CONCLUSIONS

Assets of DL-based synthesis

This Chapter is a first attempt to study the assets of DL strategies for DA in remote sensing image classification. After a proper synthesis of the source or target image to match the other acquisition, we observed an improved cross-domain portability of the classifiers. The algorithm constitutes an elegant way to align datasets, and does not depend on the dimensionality of the data sources. This last point opens interesting opportunities for cross-sensor DA, which will be explored further in future studies.

Limitations & improvements

The basic limitation of the current methodology resides in the need for labeled samples in both domains for the crucial phase where the mapping matrix is learned. While this class information allows a suitable cross-image synthesis, it also bounds the performances of the algorithm by the quality of the initial classification guess: a more promising setting would be that of a completely unsupervised synthesis. Hence, an open issue consists in finding appropriate units (instead of the land-cover classes) from which to determine the dedicated mapping function.

Part IV

CONCLUSION

10 DISCUSSION

10.1 FULFILLMENT OF THE OBJECTIVES

Original motivation

This Thesis project started with the intent to provide concrete solutions to one of the main problems currently faced in Earth observation: the difficulty in gathering ground truth samples when building (and validating) large-scale supervised land-cover classification models. In this dissertation, we addressed the issue by resorting to recently proposed developments in the field of machine learning and, more specifically, in its branch named Domain Adaptation.

Specificities of remote sensing

A mere application of these novel and highly promising techniques on a new dataset, a pair of remotely sensed images in this case, would not suffice to answer the needs of remote sensing practitioners. In fact, all the field-specific implications require a more dedicated study of the different components of the investigated methods as well as an evaluation of the best context for their application. We believe that in this Thesis the peculiarities related to the nature of the datasets we analyzed were taken into account when exploring the methods we proposed. In each Chapter of Part iii, we put forward the potential of novel Domain Adaptation methods or measures and examined their combination with more classical processing techniques already used in the remote sensing community. In order to favor the knowledge exchange between the two fields, we carefully avoided to treat the proposed adaptation procedures as black-boxes.

Fulfillment of the key objectives

Coming back to the list of specific objectives of this Thesis formulated in Section 1.2 (page 6), we can proceed with the following assessment of their fulfillment.

1. ✓ The main purpose of this work consisted in increasing the portability of the supervised classifiers across images. In this respect, we evaluated the suitability of supervised and unsupervised Domain Adaptation strategies, two means of tackling the dataset shift problem implying radically different degrees of involvements of the user. In both cases, encouraging results have been obtained on different datasets and in a range of settings. In general, the baseline of standard, non-adaptive approaches we compared them to was systematically outperformed.

2. ✓ Finding appropriate tools to evaluate the dataset shift occurring in remote sensing images acquired under different conditions was the second objective. The analyzed kernel-based measure of distance between probability distributions derived from the field of machine learning proved potential in detecting this shift and highlighting its

Table 10.1: Summary of the approaches investigated in the Thesis with their relation to the DA families and types of learning problems outlined in Chapter 4.

Approach	Principle	DA family	Learning problem	Where?
Adaptive Active Learning:				
Active Learning	Intelligently collect samples in the target domain.	Instance-transfer	Supervised DA	Chapter 6
Instance reweighting	Differently reweight source and target samples to use in target.	Instance-transfer		
Kernel-based Feature Extraction	Reduce the divergence between domains by projecting them into a new subspace.	Feature-repres.-transfer	Unsupervised DA	Chapter 7
Classic radiometric norm.:				
Histogram Matching	Relative band-by-band matching of the CDFs.	Feature-repres.-transfer	Unsupervised DA	Chapter 8
Atmospheric compensation	Absolute conversion to surface reflectance.	Feature-repres.-transfer		
Cross-image synthesis with dictionaries	Reduce the divergence between domains by synthesizing pixels via sparse representations.	Feature-repres.-transfer	Supervised DA	Chapter 9

peculiarities. Meaningful considerations about the physical processes behind the change in the pixels distribution could be derived based on this indicator.

3. ≈ The last goal of this dissertation resided in the investigation of the consequences of a change in acquisition geometry among images with a joint approach exploiting the complementarity of machine learning/statistics and physics. In this respect, the quantification and understanding of the angular effects can be deemed satisfactory. On the contrary, the efforts turned out to be insufficient to completely correct the impact of these phenomena. Both in terms of the spectral distortions of the class signatures and in terms of the portability of the land-cover models, we see room for improvement in reducing negative effects such as the reflectance anisotropies observed at various scales.

10.2 COMPARISON OF THE PRESENTED APPROACHES

In this Section, we will briefly review the solutions to remote sensing adaptation problems proposed in this Thesis. More importantly, we will put them into perspective with a comparison underlining their strengths and weaknesses. Table 10.1 recapitulates these approaches and recalls their respective Domain Adaptation and machine learning contexts. At the end of the Section, Tab. 10.2 reports instead a summary of the comparison.

- SVM-BASED ADAPTIVE ACTIVE LEARNING VIA SAMPLE REWEIGHTING:

  Summary In Chapter 6 we addressed the topic of adjusting Active Learning strategies to the situation in which the target image has seen a shift in the probability distributions. The main novelty of this study consisted in uncovering and thoroughly analyzing a sample reweighting scheme implemented using a SVM classifier. The strategy has been applied in combination with the active sampling procedure in order to intelligently re-use the information on the land-cover classes coming from the initial source image.

  Strengths & weaknesses The experimental results revealed sharper accuracy increases for the learning curves associated with the proposed strategy if compared to the baselines. Although the routine guides the sampling efforts of the analyst, the latter still has to collect ground truth labels for each new target image (supervised Domain Adaptation). This factor limits the rapid application of the method at a large scale since end-user intervention, even if minimal, is constantly required. Moreover, if field campaign and image acquisition are not simultaneous, the collected reference data could prove useless in applications with dynamic ground conditions. Nonetheless, the iterative collection of target samples combined with a reduction of the influence of misleading source samples prevents negative transfer effects. This means that as soon as the new image starts to be sampled, even though the initial source training set poorly represents the target domain, the systems is able to converge to satisfactory classification performances.

- KERNEL-BASED FEATURE EXTRACTION FOR RELATIVE NORMALIZATION:

  Summary Chapter 7 was devoted to finding meaningful projections of the data reducing the distance between the domains. We investigated the Feature Extraction paradigm, in particular examining a specific semisupervised kernel-based technique developed for Domain Adaptation. The key contribution here was the detailed study of the properties of the method as well as the remote sensing scenarios in which its application allows the best cross-image knowledge transfer.

  Strengths & weaknesses The methodology falls in the category of unsupervised Domain Adaptation approaches. This means that series of new images received by the operator can be projected to the mentioned subspace and then classified with an already trained thematic classifier, opening the way for a rapid processing of multiple images. The suitability of such a solution is backed by the good quality of the final products of the cross-image classification generally observed in the experiments. However, this system allowing such a quick mapping is heavily relying on the relevance of the initial image. If in the source image the spectral signatures of the land-cover classes are distorted to a great extent, adaption can be undermined. Another drawback of the adoption of this solution resides in the loss of the physical meaning of the variables after the projection: the spectral bands are turned into arbitrary features whose interpretation and usage for problems other than classification can be difficult.

- ANGULAR DATASET SHIFT & MODEL PORTABILITY IN MULTI-ANGLE SEQUENCES:

Summary

The primal objective of Chapter 8 was to shed light on the distortions caused by a change in the geometry of the acquisitions. By analyzing sequences of images of the same area acquired in-track by the satellite, we first focused on the physical factors controlling the imaging process. A robust kernel-based measure of distance between probability distributions showed promise in assessing the dataset shift induced by such phenomena. Subsequently, the portability of supervised classifiers across the sequence has been investigated. In this respect, we observed the evolution of the classification accuracy in the angular domain and related it to the shift highlighted with the proposed statistical measure. The substantial agreement of these trends with the underlying physical phenomena confirmed the benefits of joining the efforts of the disciplines of Earth observation and machine learning. The common denominator of the analyses mentioned above was the evaluation of the radiometric normalization abilities of traditional techniques such as atmospheric compensation and Histogram Matching.

Strengths & weaknesses

Discussing now these normalization methods, it is important to note that both of them have been extensively used in remote sensing as they maintain the physical quantities conveyed by the images. Concerning Histogram Matching, as previously remarked in the discussion for the solution based on Feature Extraction, we point out that the approach strongly depends on the relatedness of the two images to be processed. Therefore, the setting of the present case study involving a single scene certainly contributed in underestimating the negative transfer issues of this univariate matching (same thematic classes with stable proportions on the ground). Conversely, a system based on atmospheric compensation is unaffected by this type of problem in the alignment phase, as the calibration is executed with respect to an absolute reference, i. e. the surface reflectance. Nevertheless, as the setting in which the cross-image classification takes place is that of unsupervised Domain Adaptation, models trained only using ground truth data from the source image could still underachieve in the target domain. The large-scale extension of such normalization strategies can be both reasonably accurate and relatively straightforward. Indeed, Histogram Matching is quickly performed and currently developed semi-automatic atmospheric compensation routines require the input of less and less prior knowledge by the user.

- CROSS-IMAGE SYNTHESIS WITH DICTIONARIES:
  In Chapter 9, we explored the framework of Dictionary Learning and assessed its potential for synthesizing pixels with more similar characteristics across images. The investigated algorithm is based on a sparse representation of the samples. It showed an encouraging performance in finding a mapping matrix to convert the data space of a given image into that of another, a transformation ultimately increasing the cross-image portability of the classifiers.

  Summary

  One of the positive aspects of this approach is that it preserves the physical meaning of the data spaces being transformed. Indeed, the projection directly converts the spectral bands of an image into those of another (e. g. keeping pixel values in surface reflectance units), and this irrespective of the number of channels. This flexibility regarding the dimension of the data spaces comes at the expense of a more strict sampling requirement. Despite the fact that the cross-image classification itself has been carried out only based on a training set from the source image, in this preliminary phase of its development the procedure still requires labels in both domains to define the projection (supervised Domain Adaptation). In this case as well, the pertinence of the source domain is key, as a harmful knowledge transfer could happen in case of an extreme dataset shift. Such a situation is however hardly reached in practice, since the algorithm has to be applied to co-registered images. Thus, in its present form, the cross-image synthesis strategy constitutes an appropriate solution to temporal map-update problems (same scene to be classified in time) but not for land-cover mapping efforts involving multiple spatially disjoint images. Nonetheless, it is a first step towards Domain Adaptation with sparse coding, a new kind of reasoning that is becoming a major current in remote sensing image classification.

  Strengths & weaknesses

Table 10.2: Comparison of the DA approaches investigated in the Thesis.

	Adaptive Active Learning	Kernel-based Feature Extraction	Classic radiometric normalizations	Cross-image synthesis with dictionaries
System exclusively uses source labels?	✗	✓	✓	✗
Handles negative transfer?	✓	✗	≈	✗
Preserves physical meaning?	✓	✗	✓	✓
Cross-sensor knowledge transfer?	✗	✗	✗	✓
Ease of application at large-scale?	≈	✓	✓	✗
Accuracy of target land-cover map?	✓	✓	≈	≈

10.3 FURTHER WORK AND CHALLENGES

The possible extensions of the developments presented in this Thesis are numerous. They are mostly targeted at answering the needs of the remote sensing community as regards the steadily increasing amount of data acquired by the sensors. Thus, in the following we briefly recall the research directions that are worth investigating further.

Discovery of new classes

- Within the Active Learning framework, the refinement of the techniques in terms of learning curves has reached a standstill both for classic and adaptive strategies. A much more challenging topic is that of discovering and handling new land-cover classes in the iterative process. Attention could be paid to approaches favoring a sampling heuristic based on a diversity criterion in the first iterations to comprehensively search the input space and then gradually turning to the more conventional class boundary refinement objective.

From cross-image to cross-sensor

- Another central aspect of adaptation that only recently started to draw the attention of the scientists in Earth observation concerns the ability to cope with images acquired by different sensors. If in change detection this line of research is more mature, when dealing with land-cover model portability much work is still needed. The fact the images are not co-located makes the definition of such a cross-sensor mapping more difficult. Although ambitious, this objective could lead to unprecedented opportunities in terms of constant and spatially extensive monitoring efforts. Not having to always rely on acquisitions by the same specific sensor will enable the user to flexibly re-use the collected ground truth, minimizing thus his/her onerous involvement in the mapping process.

- *Object-level knowledge transfer* Throughout this manuscript, all the approaches we explored had one thing in common. During both the optional projection phase and the cross-image classification step, the basic spatial unit we considered was the pixel, even if in some cases neighborhood information has been included. We believe that a key research question for the future resides in the study of adaptation strategies working at the object level, thereby replacing the traditional pixel-based strategies.

- *Angular effects* A complete understanding of the angular effects impacting acquisitions bearing different view angles is also still an open issue. In this Thesis we made an attempt in this direction but, to paint the full picture, more analyses are definitely required. For instance, small-scale anisotropic reflectance behaviors that greatly modify the spectral signature of certain materials, the BRDF effects, have yet to be specifically investigated and properly compensated. To this end, auspicious results can be expected with scientific studies at the interface of statistics and physics, two complementary disciplines playing a central role in the development of the remote sensing technology.

- *Truly large-scale applications* The solutions proposed in this Thesis, as well as most of the works proposed in the literature are actually tested on image subsets that are often orders of magnitude smaller than the original acquisitions collected by the sensors (e. g. a WorldView-2 panchromatic image generally has a size of more than 30,000×30,000 pixels). In this context, even when working with single images, the sample selection bias arising from small sampling regions has not to be underestimated. Therefore, to pursue studies of the land-cover at a truly large spatial scale, more development and validation efforts for the adaptation methodologies are clearly needed. With this objective in mind, we believe complex approaches should be avoided in concrete applications, giving the priority to simpler solutions (e. g. linear models, classic kernel-based classifiers, basic compensation strategies).

Part V

APPENDIX

A

CLASSIFICATION QUALITY MEASURES

Supervised classification assessment

Considering the special case of a classification task involving remote sensing data, in this Appendix we present some useful measures taken as the gold standard by the community when assessing the quality of the thematic maps produced with a supervised classifier [Foody, 2002, 2004].

Confusion matrix

The starting point is the confusion matrix, which is the result of a cross-tabulation of actual (observed ground truth) and predicted (by the classifier) classes. This matrix that allows to subsequently derive the quality measures is outlined in Tab. A.1.

Table A.1: Confusion matrix for a multi-class prediction with c classes concerning $n_{\bullet\bullet}$ samples. PA: Producer's Accuracy, UA: User's Accuracy.

		Actual class					
		1	2	$\cdots$	c	Totals	UA
Predicted class	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1\bullet}$	$n_{11}/n_{1\bullet}$
	2	n_{21}	n_{22}			$n_{2\bullet}$	$n_{22}/n_{2\bullet}$
	$\vdots$	$\vdots$		$\ddots$		$\vdots$	$\vdots$
	c	n_{c1}			n_{cc}	$n_{c\bullet}$	$n_{cc}/n_{c\bullet}$
Totals		$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet c}$	$n_{\bullet\bullet}$	
PA		$n_{11}/n_{\bullet 1}$	$n_{22}/n_{\bullet 2}$	$\cdots$	$n_{cc}/n_{\bullet c}$		

A.1 OVERALL MEASURES

Overall Accuracy

The most common measure is the *Overall Accuracy* (OA) (ranging in $[0, 1]$ with best score 1), which is the sum of pixels correctly classified in each class, n_{ii}, divided by the total number of pixels involved in the prediction, $n_{\bullet\bullet}$:

$$\mathrm{OA} = \frac{\sum_{i=1}^{c} n_{ii}}{n_{\bullet\bullet}} \ . \tag{A.1}$$

Kappa statistic

The *Kappa statistic* κ (ranging in $[-1, 1]$ with best score 1), also referred to as Cohen's Kappa coefficient of agreement [Cohen, 1960], provides a more complete measure of the accuracy of the prediction. Indeed, contrary to the previously presented OA which only considers the information in the diagonal of the confusion matrix, this index makes use of the entries of the whole table. An estimate of Kappa is provided by

$$\kappa = \frac{p_\mathrm{o} - p_\mathrm{c}}{1 - p_\mathrm{c}} \ , \tag{A.2}$$

where $p_o = \sum_i n_{ii}/n_{\bullet\bullet}$ is the observed proportion of correctly classified pixels, i. e. the OA, and $p_c = \sum_i n_{i\bullet}n_{\bullet i}/n^2_{\bullet\bullet}$ is the proportion of correctly classified pixels that is expected by chance. This metric may be interpreted as a measure of the improvement ensured by the classifier at hand over a random allocation the predicted labels. Moreover, when dealing with a large imbalance in the actual class counts, Kappa is much better suited, with respect to OA, to provide an unbiased measure of accuracy also appropriately factoring in the errors committed on small classes.

McNemar's test

Finally, when the test site where the ground truth has been collected is the same, a direct comparison between two classifiers can be carried out via a McNemar's test [Bradley, 1968]. This is a non-parametric test that is based on the following standardized normal test statistic [Foody, 2004]:

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \ . \tag{A.3}$$

The quantity n_{12} indicates the number of pixels correctly predicted by classifier 1 while simultaneously being incorrectly predicted by classifier 2. Conversely, n_{21} represents the number of samples incorrectly predicted by classifier 1 and correctly predicted by classifier 2. Under the null hypothesis H_0 stating that the two classifiers are equivalent ($n_{12} = n_{21}$), the McNemar's test z value follows a normal distribution. Thus, running a two-tailed test with the standard α level of 5%, a value $z > 1.96$ indicates a statistically significant superiority of classifier 1 over classifier 2.

A.2 CLASS-SPECIFIC MEASURES

User's and Producer's Accuracy

The associated individual class accuracies are named *User's Accuracy* (UA) and *Producer's Accuracy* (PA) (see last column and row of Tab A.1, respectively). On the one hand, UA accounts for the *commission error* (1 − UA), i. e. the proportion of predicted pixels wrongly allocated to a given class by the model. UA provides map users with accuracy information indicating the quality of the thematic map. In fact, it is nothing but the probability that a pixel classified in a given class actually represents that same class on the ground. On the other hand, PA is complementary to the *omission error* (1 − PA), that is the proportion of ground truth pixels wrongly assigned to other classes. PA helps the map producer to evaluate and refine the mapping (prediction) process, as this measure denotes the probability that an actual ground truth pixel has been correctly classified by the model.

F-measure

An efficient way to provide a single class-specific indicator combining UA and PA is represented by the *F-measure*, which is computed as follows

$$\text{F-measure} = 2 \cdot \frac{\text{UA} \cdot \text{PA}}{\text{UA} + \text{PA}} \ . \tag{A.4}$$

Such a statistic, the harmonic mean of the two class-specific measures, is usually employed in information retrieval as a means to combine *precision* and *recall* [Powers, 2011], the equivalents of UA and PA in binary classification problems.

B

DATASETS USED IN THE THESIS

This Appendix describes the 5 remote sensing datasets that have been used in the Thesis:

- B.1 QuickBird images of Zurich
- B.2 AVIRIS images of the Kennedy Space Center
- B.3 ROSIS image of Pavia
- B.4 WorldView-2 multi-angle sequence of Atlanta
- B.5 WorldView-2 multi-angle sequence of Rio de Janeiro

We would like to thank Prof. Melba Crawford at Purdue University for making the Kennedy Space Center AVIRIS data available. Prof. Paolo Gamba from the University of Pavia is acknowledged for providing the Pavia ROSIS image. We also thank Dr. Nathan Longbotham and Dr. Fabio Pacifici at DigitalGlobe for supplying and preprocessing the Atlanta and Rio de Janeiro WorldView-2 imagery.

The Pavia ROSIS dataset and part of the Rio de Janeiro WorldView-2 sequence (5 angular acquisitions) have been distributed in the context of the IEEE GRSS Data Fusion Contest of 2008 and 2011, respectively. In this regard, we acknowledge the GRSS Image Analysis and Data Fusion Technical Committee.

B.1 QUICKBIRD IMAGES OF ZURICH

Acquisition and size

The Zurich dataset consists of two VHR QuickBird images of the city of Zurich (Switzerland), representing two spatially distant neighborhoods. The target image is a 474×482 pixels subset of an acquisition of August 2002 while the source image is a 301×296 pixels subset of an acquisition of October 2006. Figures B.1(a) and (c) illustrate the two considered images.

Bands and spatial resolution

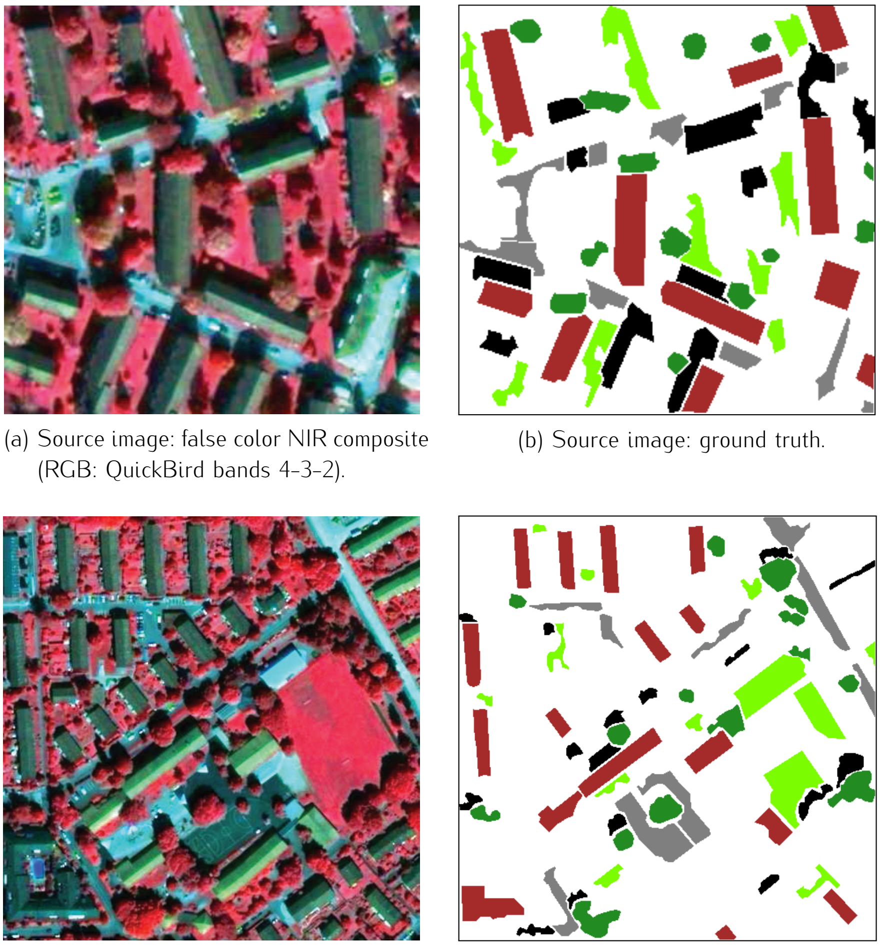
The images present 4 multispectral VNIR bands and a PAN band covering the region of the spectrum from 450 to 900 nm (see Tab. 2.1 on page 18). The multispectral bands, originally possessing a spatial resolution of 2.4 m, have been pansharpened with the Gram-Schmidt method [Brower and Laben, 2000] to reach a spatial resolution of 0.6 m.

Land-cover classes

The ground truth defined by visual inspection includes pixels from 4 land-cover classes characterizing both images: "buildings", "roads", "grass", "vegetation". An extra class "shadows" has been arbitrarily added, bringing the total of classes to 5. Figures B.1(b) and (d) show the ground truth maps for the source and target images, respectively. Table B.1 details the class counts per image and presents a legend of the colors used in the maps.

Reasons of the dataset shift

The differences in marginal and class-conditional distributions between the source and the target image are caused by three factors: 1) differences in illumination conditions (sun and satellite elevations have changed), 2) seasonal effects affecting vegetation growth and 3) varying materials composing roofs and roads.



(a) Source image: false color NIR composite (RGB: QuickBird bands 4-3-2).

(b) Source image: ground truth.

(c) Target image: false color NIR composite (RGB: QuickBird bands 4-3-2).

(d) Target image: ground truth.

Figure B.1: QuickBird images of the city of Zurich. The source image is of size 301×296 pixels. The target image is of size 474×482 pixels. Ground truth: 5 thematic classes. For the legend refer to Tab. B.1.

Table B.1: Zurich dataset: names, number of labeled pixels per image and colors for the land-cover classes.

Class name	# source samples	# target samples	Color
buildings	10,729	15,897	🟥
roads	4,970	10,050	◼
shadows	3,159	8,551	⬛
trees	3,324	9,981	🟩
grass	6,062	5,041	🟩
TOTAL	28,244	49,520	

B.2 AVIRIS IMAGES OF THE KENNEDY SPACE CENTER

Acquisition and size

The KSC dataset comprises two sub-regions of the same hyperspectral acquisition that has been obtained over the Kennedy Space Center, Florida (USA), on March 23 1996 [Rajan et al., 2008]. Figures B.2(a) and (c) illustrate these two 614×512 pixels subsets.

Bands and spatial resolution

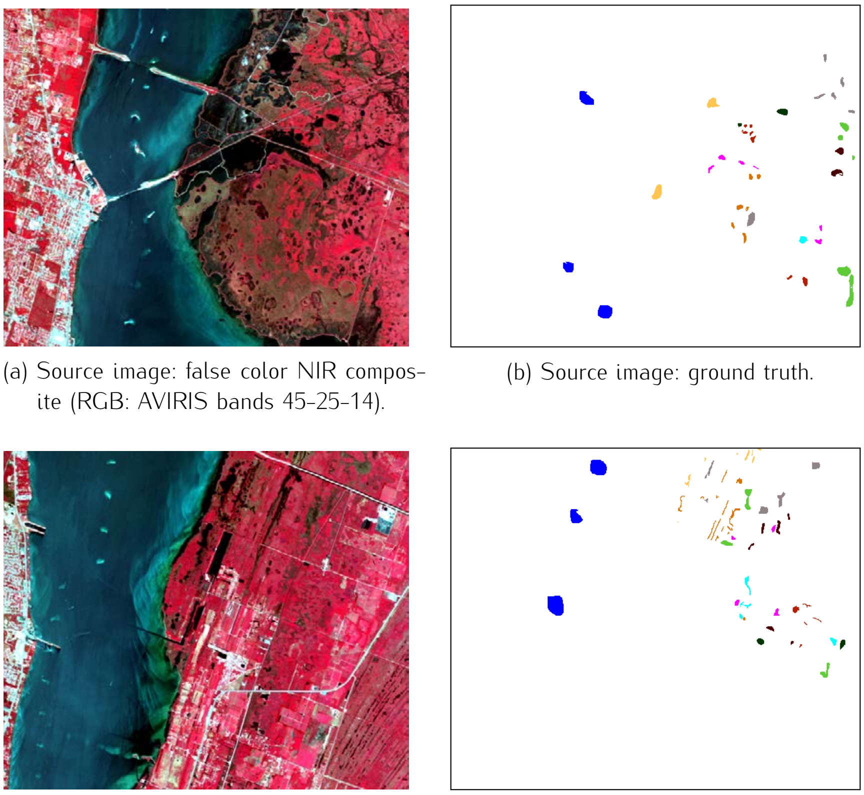
The image has been acquired with the airborne AVIRIS hyperspectral instrument and counts 224 bands covering the region between 400 and 2500 nm. After the removal of water absorption and low SNR bands, the dataset is composed of a total of 176 bands (indices of the original bands kept: 5-101, 117-150, 173-217). The spatial resolution of the image is 18 m.

Land-cover classes

The retained ground truth only includes land-cover classes that are found in both images. Figures B.2(b) and (d) depict the ground truth maps for the source and target images, respectively. The list of these 10 classes, mainly consisting of types of subtropical vegetation, along with details about the class counts per image and a legend of the colors used in the maps is given in Tab. B.2.

Reasons of the dataset shift

The spectra of the classes present a rather large variation across the two retained areas, justifying the definition of distinct source and target domains. Some of the classes have been defined as mixed land-covers. Therefore, slight changes in the proportions of these end-members throughout the image will cause a shift in the probability distributions.



(a) Source image: false color NIR composite (RGB: AVIRIS bands 45-25-14).

(b) Source image: ground truth.

(c) Target image: false color NIR composite (RGB: AVIRIS bands 45-25-14).

(d) Target image: ground truth.

Figure B.2: AVIRIS images of the Kennedy Space Center. The both sub-regions are of size 614×512 pixels. Ground truth: 10 thematic classes. For the legend refer to Tab. B.2.

Table B.2: KSC dataset: names, number of labeled pixels per image and colors for the land-cover classes.

Class name	# source samples	# target samples	Color
scrub	761	422	🟩
willow swamp	243	180	🟪
cabbage palm hammock	256	431	🟧
cabbage palm/oak hammock	252	132	🟥
slash pine	161	166	🟩
oak/broadleaf hammock	229	274	🟫
hardwood swamp	105	248	🟦
graminoid marsh	431	453	▨
salt marsh	419	156	🟨
water	927	1,392	🟦
TOTAL	3,784	3,854	

B.3 ROSIS IMAGE OF PAVIA

Acquisition and size

The Pavia dataset consists of a single image acquired by the airborne ROSIS-03 hyperspectral sensor with a flight over the city center of Pavia (Italy) operated by DLR [Licciardi et al., 2009] (see Fig. B.3(a)). From the original 1400×512 scene, a source sub-region was defined on a patch of 172×123 pixels whereas a target sub-region was set to cover a separate and larger 350×350 pixels area.

Bands and spatial resolution

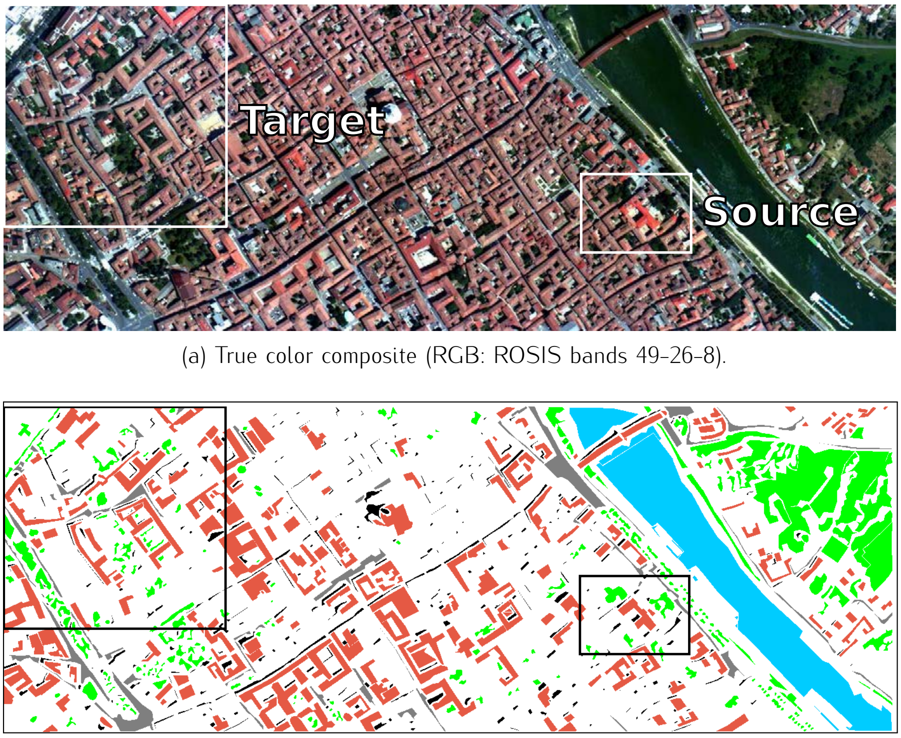
Although the sensor acquires a total of 115 spectral bands covering a region of the spectrum between 430 and 860 nm, due to the presence of 13 noisy channels, only 102 bands were retained for the analyses. The associated spatial resolution is 1.3 m.

Land-cover classes

The captured scene is mainly representing an urban setting: 4 thematic classes have been delineated throughout the image: "buildings", "roads", "shadows" and "vegetation". Figure B.3(b) shows the ground truth map. Details about the labeled samples located in the source and target sub-regions are given in Tab. B.3. Note that the original dataset also includes the thematic class "water", excluded here as not present across the entire image.

Reasons of the dataset shift

The different nature of the materials constituting roofs and roads as well as the presence of various types of vegetation, cause a remarkable variation across the image of the spectral signatures of these land-cover classes. In this context, we could consider the two disjoint subsets of the scene as two separate domains.



(a) True color composite (RGB: ROSIS bands 49-26-8).

(b) Ground truth: 5 thematic classes (only 4 were used in the adaptation experiments as "water" was excluded). For the legend refer to Tab. B.3.

Figure B.3: ROSIS image of the city center of Pavia. The sub-regions considered as source (172×123 pixels patch on the right) and target (350×350 pixels patch on the left) images are indicated with white polygons in (a) and with black polygons in (b).

Table B.3: Pavia dataset: names, number of labeled pixels per image sub-region and colors for the land-cover classes.

Class name	# source samples	# target samples	Color
buildings	1,465	17,501	🟥
roads	326	2,549	▨
shadows	514	1,638	⬛
vegetation	1,793	6,406	🟩
water	-	-	🟦
TOTAL	4,098	28,094	

B.4 WORLDVIEW-2 MULTI-ANGLE SEQUENCE OF ATLANTA

Acquisition, geometry of the collection and size

The Atlanta dataset consists of a multi-angle in-track sequence collected by WorldView-2 during a 2-minute time frame over the city center of Atlanta, Georgia (USA), in December 2009. From this multi-angular acquisition, we retained 13 images with an off-nadir angle varying from −33.2° (i. e. 33.2°, looking southward from the satellite to the imaged area) to +31.5° (i. e. 31.5°, looking northward from the satellite to the imaged area). Each image covers exactly the same area and has a size of 1907×1266 pixels (a subset of the complete scene). Figure B.4(a) illustrates the most nadiral acquisition of the sequence, while Fig. B.5 reports the satellite elevations and azimuths along the collection track.

Bands and spatial resolution

The images present 8 multispectral bands between 400 nm and 1050 nm. In fact, WorldView-2 extends with 4 additional spectral channels (coastal, yellow, red edge, NIR2) the standard 4 of the QuickBird sensor (blue, green, red, NIR). The spatial resolution is 2 meters (see Tab. 2.1 on page 18).

Land-cover classes

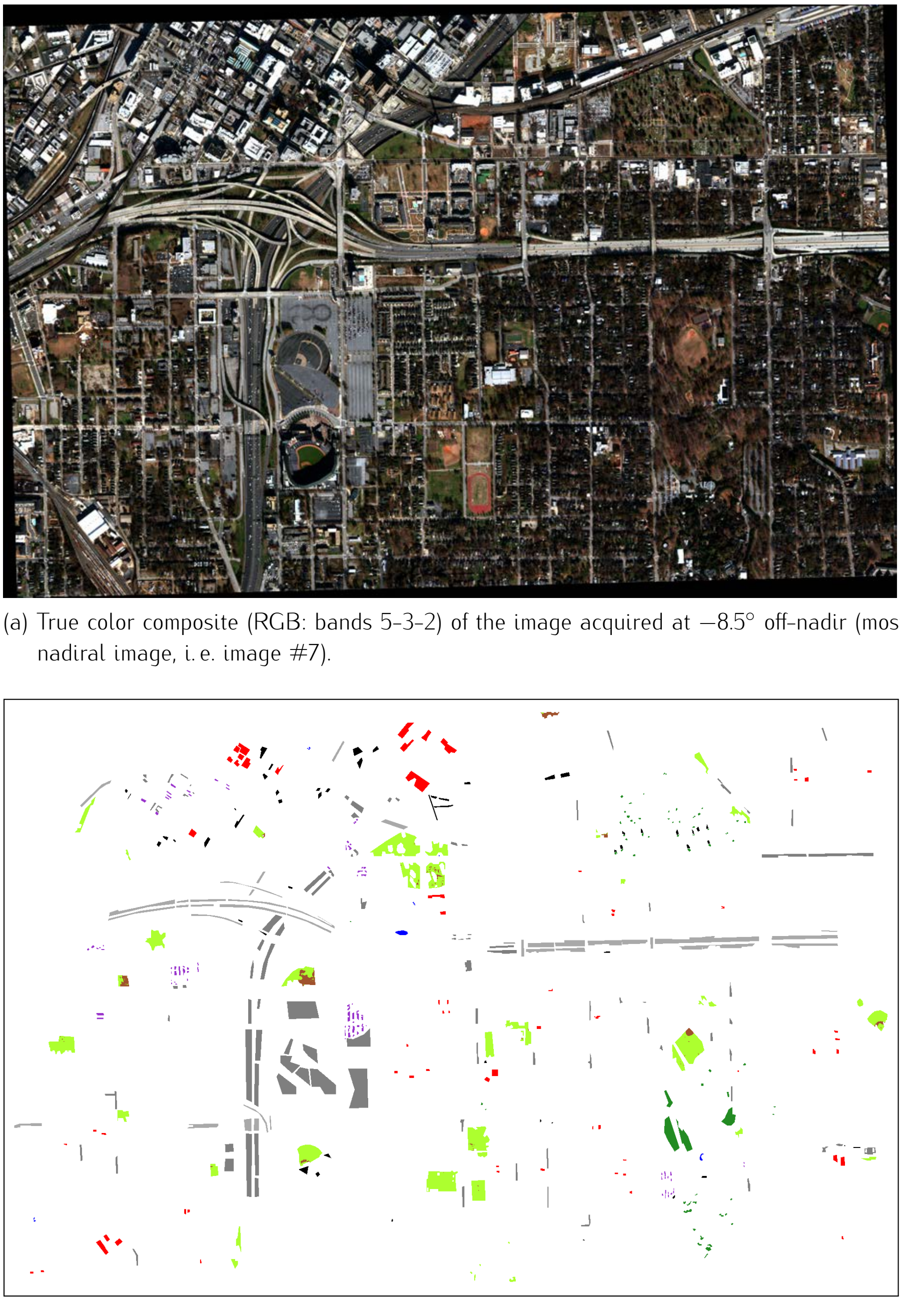
Since the sequence has been collected over part of downtown Atlanta, a set of classes commonly found in urban environments has been considered. Hence, the ground truth included different kinds of vegetative cover, several types of man-made objects and urban structures found across the entire scene. The 9 ground-cover classes ultimately identified are listed in Tab. B.4 with the respective class counts and map legend colors.

Validitiy of the ground truth

The reference polygons could be propagated through the whole multi-angular sequence by using true-orthorectified images obtained thanks to a *Digital Surface Model* [Longbotham et al., 2012a]. However, as abrupt changes in elevation (e. g. high buildings) could produce occlusion artifacts, the survey has been carried out to collect samples in open areas with a relatively small topographic variation. Figure B.4(b) reports the common ground truth map valid for all the sequence.

Reasons of the dataset shift

Three main factors cause the observed angular dataset shift in this sequence. First, we remark an increased Rayleigh scattering at high off-nadir angles, yielding hazy images in these angular regions of the sequence. Second, small-scale BRDF effects are clearly visible for some specific surfaces (asphalt, grass, etc.). Third, at a larger object scale, the solar observational cross-section causes remarkably different scattering/shadowing behavior along the path. In this regard, starting from the northernmost acquisition, note that images #1 (−33.2°) to #8 (+9.5°) lie in the forward solar scattering region. This means they have been acquired opposite the sun with respect to the target area. The remaining images #9 (+12.2°) to #13 (+31.5°) are instead in the backward scattering region, where the sun and the satellite look at the imaged area from the same side (see Fig. B.5).



(a) True color composite (RGB: bands 5-3-2) of the image acquired at −8.5° off-nadir (most nadiral image, i. e. image #7).

(b) Ground truth: 9 thematic classes. For the legend refer to Tab. B.4.

Figure B.4: WorldView-2 images of the city center of Atlanta.

Table B.4: Atlanta dataset: names, number of labeled pixels and colors for the land-cover classes.

Class name	# samples	Color
water	313	🟦
concrete	16,479	⬜
asphalt	30,099	◼
vehicles	1,759	🟪
soil	2,014	🟫
grass	27,561	🟩
buildings	6,283	🟥
shadow	3,156	⬛
trees	3,905	🟩
TOTAL	91,569	

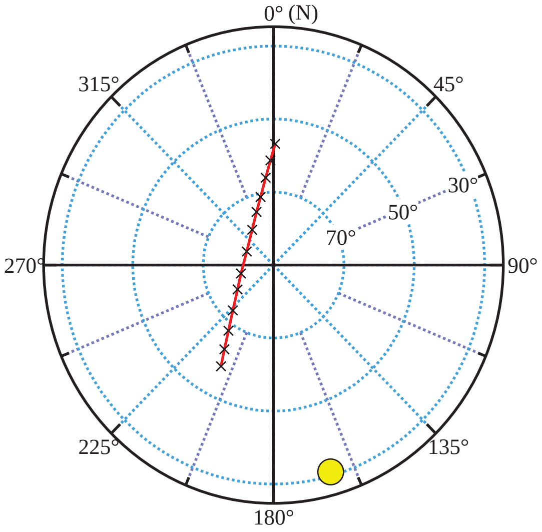


Figure B.5: Atlanta dataset: ground observed azimuth (plotted angularly clockwise: north = 0°, east = 90°, south = 180°, west = 270°) and elevation (plotted radially from the center: ground nadir = 90°, ground horizon = 0°) of the satellite for each acquisition in the sequence (black crosses) as well as for the sun (yellow circle). Images are identified as #1 (−33.2° off-nadir) to #13 (+31.5° off-nadir), starting from the northernmost acquisition.

B.5 WORLDVIEW-2 MULTI-ANGLE SEQUENCE OF RIO DE JANEIRO

Acquisition, geometry of the collection and size

The Rio de Janeiro dataset is an angular in-track sequence acquired by WorldView-2 over of the city center of Rio de Janeiro (Brazil), in January 2010. The sequence, obtained during a 5-minute collection period, consists of 20 images with off-nadir angles going from −47.3° (i. e. 47.3°, looking southward from the satellite to the imaged area) to +47.5° (i. e. 47.5°, looking northward from the satellite to the imaged area). The considered scene is the same across all the acquisitions and is a subset of size 463×328 pixels of the imaged area. Figure B.6(a) pictures the most nadiral image of the sequence. Figure B.7 shows the satellite elevations and related azimuths along the collection track.

Bands and spatial resolution

The images possess the same spectral channels and the same associated spatial resolution of the Atlanta dataset presented in Appendix B.4: 8 bands (400 nm to 1050 nm) at a spatial resolution of 2 m.

Land-cover classes

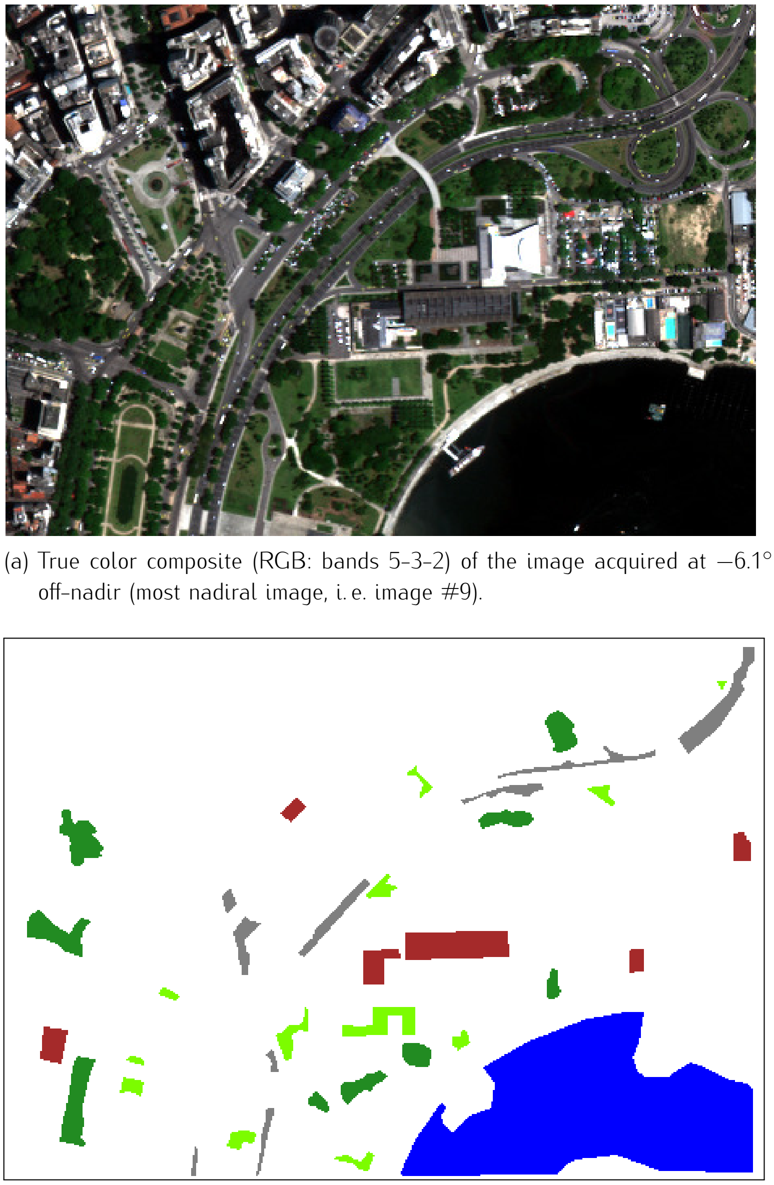
The imaged scene concerns an area just south of downtown Rio de Janeiro. We observe several large buildings, roads of varying size, community parks as well as part of the bay. Table B.5 details the 5 land-cover classes, and their colors used in the maps, which have been manually delineated on the images by photo-interpretation. The corresponding class counts refer to the image acquired at −6.1° off-nadir and can be considered as representative for the entire sequence.

Validitiy of the ground truth

In fact, for this dataset, the images were not true-orthorectified, i. e. there was not a perfect pixel-by-pixel superimposition throughout the sequence. This required us to provide a separate ground truth for each acquisition, though always including the same objects. Figure B.6(b) reports the ground truth map for the image acquired at −6.1° off-nadir, the most nadiral acquisition.

Reasons of the dataset shift

As for the Atlanta dataset, the geometry of the acquisition and the related angular effects are the only factors inducing the probability shift for these images. However, the solar observational cross-section effects (third factor) have changed. If compared to the Atlanta sequence, the Rio de Janeiro acquisition took place with a different combination of satellite-sun positions. In this case, the sun was almost perpendicular to the satellite flight path, causing a more symmetrical scattering/shadowing behavior along the sequence. No clear distinction between forward or backward scattering regimes can be made.



(a) True color composite (RGB: bands 5-3-2) of the image acquired at −6.1° off-nadir (most nadiral image, i. e. image #9).

(b) Ground truth: 5 thematic classes. It refers to the image acquired at −6.1° off-nadir (most nadiral image, i. e. image #9). For the legend refer to Tab. B.5.

Figure B.6: WorldView-2 images of the city center of Rio de Janeiro.

Table B.5: Rio de Janeiro dataset: names, number of labeled pixels and colors for the land-cover classes (relative to the ground truth of the image acquired at −6.1° off-nadir).

Class name	# samples	Color
water	13,532	🟦
grass	1,564	🟩
roads	2,047	◼
trees	2,946	🟩
buildings	2,042	🟥
TOTAL	22,131	

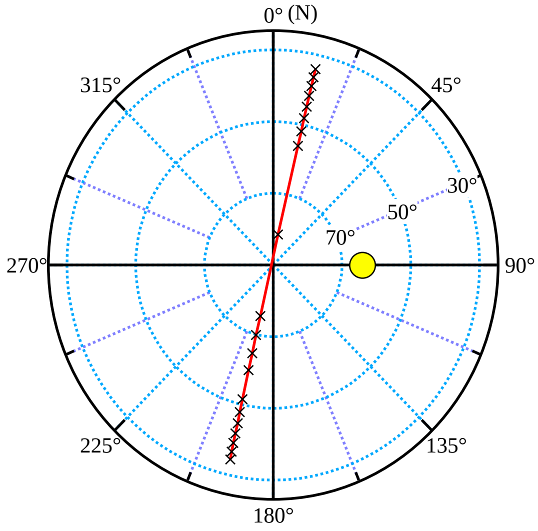


Figure B.7: Rio de Janeiro dataset: ground observed azimuth (plotted angularly clockwise: north = 0°, east = 90°, south = 180°, west = 270°) and elevation (plotted radially from the center: ground nadir = 90°, ground horizon = 0°) of the satellite for each acquisition in the sequence (black crosses) as well as for the sun (yellow circle). Images are identified as #1 (−47.3° off-nadir) to #20 (+47.6° off-nadir), starting from the northernmost acquisition.

BIBLIOGRAPHY

F. Achard, H. D. Eva, H.-J. Stibig, P. Mayaux, J. Gallego, T. Richards, and J.-P. Malingreau. Determination of deforestation rates of the world's humid tropical forests. *Science*, 297 (5583):999–1002, 2002. (Cited on page 12.)

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. In *Proceedings of the Signal Processing with Adaptive Sparse Structured Representations workshop (SPARS)*, volume 5, pages 9–12, 2005. (Cited on pages 43 and 131.)

N. Alajlan, E. Pasolli, F. Melgani, and A. Franzoso. Large-scale image classification using active learning. *IEEE Geoscience and Remote Sensing Letters*, 11(1):259–263, 2014. (Cited on page 64.)

J. Arenas-García and K. B. Petersen. Kernel multivariate analysis in remote sensing feature extraction. In G. Camps-Valls and L. Bruzzone, editors, *Kernel Methods for Remote Sensing Data Analysis*. J. Wiley & Sons, NJ, USA, 2009. (Cited on pages 41 and 65.)

G. P. Asner. Biophysical and biochemical sources of variability in canopy reflectance. *Remote Sensing of Environment*, 64(3):234–253, 1998. (Cited on page 12.)

S. K. Babey and C. D. Anger. Compact airborne spectrographic imager (CASI): a progress review. In *Proceeding of the SPIE conference on Imaging Spectrometry of the Terrestrial Environment*, pages 152–163, 1993. (Cited on page 18.)

C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. Exploiting manifold geometry in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):441–454, 2005. (Cited on page 65.)

K. Bahirat, F. Bovolo, L. Bruzzone, and S. Chaudhuri. A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains. *IEEE Transactions on Geoscience and Remote Sensing*, 50(7):2810–2826, 2012. (Cited on page 67.)

M. J. Barnsley, J. J. Settle, M. A. Cutter, D. R. Lobb, and F. Teston. The PROBA/CHRIS mission: A low-cost smallsat for hyperspectral multiangle observations of the Earth surface and atmosphere. *IEEE Transactions on Geoscience and Remote Sensing*, 42(7):1512–1520, 2004. (Cited on page 18.)

G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000. (Cited on page 40.)

Y. Bazi and F. Melgani. Gaussian process approach to remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1):186–197, 2010. (Cited on page 20.)

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006. (Cited on page 90.)

J. A. Benediktsson, P. H. Swain, and O. K. Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):540–552, 1990. (Cited on page 20.)

A. Berk, L. S. Bernstein, and D. C. Robertson. MODTRAN: A moderate resolution model for LOWTRAN. Technical report, DTIC Document, 1987. (Cited on page 24.)

A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35: 99–109, 1943. (Cited on page 50.)

J. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, 2013. (Cited on page 19.)

C. M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006. (Cited on page 29.)

J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics, 2006. (Cited on page 56.)

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Association For Computational Linguistics*, pages 440–447, 2007. (Cited on page 56.)

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. (Cited on pages 51 and 68.)

J. V. Bradley. *Distribution-free statistical tests*. NJ, Prentice-Hall, 1968. (Cited on page 150.)

C. Brekke and A. H. S. Solberg. Oil spill detection by satellite remote sensing. *Remote Sensing of Environment*, 95(1):1–13, 2005. (Cited on page 12.)

B. V. Brower and C. A. Laben. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening, January 4 2000. US Patent 6,011,875. (Cited on pages 17 and 152.)

L. Bruzzone and R. Cossu. A multiple-cascade-classifier system for a robust and partially unsupervised updating of land-cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):1984–1996, 2002. (Cited on page 67.)

L. Bruzzone and R. Cossu. An adaptive approach to reducing registration noise effects in unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2455–2465, 2003. (Cited on page 20.)

L. Bruzzone and D. Fernàndez Prieto. Unsupervised retraining of a maximum likelihood classifier for the analysis of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2):456–460, 2001. (Cited on page 67.)

L. Bruzzone and M. Marconcini. Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 47(4):1108–1122, 2009. (Cited on page 67.)

L. Bruzzone and M. Marconcini. Domain adaptation problems: a DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010. (Cited on pages 22 and 58.)

L. Bruzzone and C. Persello. A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability. *IEEE Transactions on Geoscience and Remote Sensing*, 47(9):3180–3191, 2009. (Cited on pages 50 and 66.)

L. Bruzzone and D. F. Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 38(3):1171–1182, 2000. (Cited on page 20.)

L. Bruzzone, F. Roli, and S. B. Serpico. An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, 33(6):1318–1321, 1995. (Cited on page 65.)

L. Bruzzone, M. Chi, and M. Marconcini. A Novel Transductive SVM for Semisupervised Classification of Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11):3363–3373, 2006. (Cited on page 20.)

R. Caloz and C. Collet. *Précis de télédétection; vol. 3: Traitements numériques d'images de télédétection*. Presses de l'Université de Québec, Agence universitaire de la Francophonie., 2001. (Cited on page 19.)

G. Camps-Valls. Machine learning in remote sensing data processing. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6, 2009. (Cited on page 30.)

G. Camps-Valls and L. Bruzzone. Kernel-based methods for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 43(6):1351–1362, 2005. (Cited on pages 20 and 38.)

G. Camps-Valls, T. V. Bandos, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3044–3054, 2007. (Cited on page 20.)

G. Camps-Valls, J. Mooij, and B. Scholkopf. Remote sensing feature selection by kernel dependence measures. *IEEE Geoscience and Remote Sensing Letters*, 7(3):587–591, 2010. (Cited on pages 65 and 90.)

G. Camps-Valls, J. Malo, D. Tuia, and L. Gomez-Chova. *Remote sensing image processing*, volume 12. Morgan & Claypool Publishers, 2011. (Cited on page 30.)

G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson. Advances in hyperspectral image classification: Earth monitoring with statistical learning methods. *IEEE Signal Processing Magazine*, 31:45–54, 2014. (Cited on page 30.)

M. J. Canty and A. A. Nielsen. Automatic radiometric normalization of multitemporal satellite imagery with the iteratively re-weighted mad transformation. *Remote Sensing of Environment*, 112(3):1025–1036, 2008. (Cited on page 65.)

M. J. Canty, A. A. Nielsen, and M. Schmidt. Automatic radiometric normalization of multitemporal satellite imagery. *Remote Sensing of Environment*, 91(3-4):441–451, 2004. (Cited on pages 24 and 65.)

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. (Cited on page 76.)

M. W. Chang, C. J. Lin, and R. C. Weng. Analysis of switching dynamics with competing support vector machines. *IEEE Transactions on Neural Networks*, 15:720–727, 2004. (Cited on page 70.)

P. S. Jr. Chavez. An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment*, 24(3):459–479, 1988. (Cited on page 23.)

M. Chen, K. Q. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2456–2464, 2011a. (Cited on page 56.)

Y. Chen, N. M. Nasrabadi, and T. D. Tran. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 49 (10):3973–3985, 2011b. (Cited on page 44.)

V. Cherkassky and F. Mulier. *Learning from Data - Concepts, Theory and Methods*. Wiley, 2007. (Cited on pages 29 and 31.)

R. N. Clark and T. L. Roush. Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research, B: Solid Earth (1978–2012)*, 89(B7):6329–6340, 1984. (Cited on page 16.)

T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields. The HyMapTM airborne hyperspectral sensor: the system, calibration and performance. In *1st EARSeL workshop on imaging spectrometry*, pages 37–42, 1998. (Cited on page 18.)

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960. (Cited on page 149.)

T. Cooley, G. P. Anderson, G. W. Felde, M. L. Hoke, A. J. Ratkowski, J. H. Chetwynd, J. A. Gardner, S. M. Adler-Golden, M. W. Matthew, A. Berk, L. S. Bernstein, P. K. Acharya, D. Miller, and P. Lewis. FLAASH, a MODTRAN4-based atmospheric correction algorithm, its application and validation. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1414–1418, 2002. (Cited on page 23.)

P. R. Coppin and M. E. Bauer. Digital change detection in forest ecosystems with remote sensing imagery. *Remote Sensing Reviews*, 13(3-4):207–234, 1996. (Cited on page 21.)

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, (3):

326–334, 1965. (Cited on page 37.)

M. M. Crawford, D. Tuia, and H. L. Yang. Active learning: Any value for classification of remotely sensed data? *Proceedings of the IEEE*, 101(3):593–608, 2013. (Cited on pages 21 and 64.)

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000. (Cited on pages 36 and 37.)

J. C. Curlander and R. N. McDonough. *Synthetic aperture radar*, volume 199. Wiley New York, 1991. (Cited on page 11.)

W. Dai, Q. Yang, G.-R. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 193–200, New York, NY, USA, 2007. (Cited on pages 54, 64, 69, 71, and 72.)

M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone. Morphological attribute profiles for the analysis of very high resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 48(10):3747–3762, 2010. (Cited on page 21.)

H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the Annual Meeting of the Association For Computational Linguistics*, pages 256–263, Prague, 2007. (Cited on page 56.)

H. Daumé III, A. Kumar, and A. Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59. Association for Computational Linguistics, 2010. (Cited on page 56.)

S. M. Davis, D. A. Landgrebe, T. L. Phillips, P. H. Swain, R. M. Hoffer, J. C. Lindenlaub, and L. F. Silva. *Remote sensing: the quantitative approach*, volume 1. McGraw-Hill International Book Co., New York, 405 p., 1978. (Cited on page 30.)

B. Demir, C. Persello, and L. Bruzzone. Batch-mode active-learning methods for the interactive classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 49(3):1014–1031, 2011. (Cited on page 64.)

W. Di and M. M. Crawford. View generation for multiview maximum disagreement based active learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5):1942–1954, 2012. (Cited on page 64.)

D. J. Diner, J. C. Beckert, T. H. Reilly, C. J. Bruegge, J. E. Conel, R. A. Kahn, J. V. Martonchik, T. P. Ackerman, R. Davies, S. A. W. Gerstl, H. R. Gordon, Muller J.-P., Myneni R., Sellers P. J., Pinty B., and Verstraete M. M. Multi-angle imaging spectroradiometer (MISR) instrument description and experiment overview. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4):1072–1087, 1998. (Cited on page 19.)

L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer SVM for video concept detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1375–1381, 2009. (Cited on page 58.)

L. Duan, I. W. Tsang, and D. Xu. Domain transfer multiple kernel learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):465–479, 2012. (Cited on pages 58 and 68.)

R. Duca and F. Del Frate. Hyperspectral and multiangle CHRIS–PROBA images for the generation of land cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, 46(10):2857–2866, 2008. (Cited on page 113.)

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, 2001. (Cited on pages 31 and 40.)

E. Eaton and M. desJardins. Set-based boosting for instance-level transfer. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 422–428, 2009. (Cited on page 54.)

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7:179–188, 1936. (Cited on page 39.)

M. D. Fleming, J. S. Berkebile, and R. M. Hoffer. Computer-aided analysis of LANDSAT-I MSS data: a comparison of three approaches, including a "Modified clustering" approach.

Technical report, LARS information note 072475 Purdue University, 1975. (Cited on page 63.)

M. A. Folkman, J. Pearlman, L. Liao, and P. J. Jarecke. EO-1/Hyperion hyperspectral imager design, development, characterization, and calibration. In *Proceeding of the SPIE conference on Hyperspectral Remote Sensing of the Land and Atmosphere*, volume 4151, pages 40–51, 2001. (Cited on page 18.)

G. M. Foody. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80(1):185–201, 2002. (Cited on page 149.)

G. M. Foody. Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy. *Photogrammetric Engineering and Remote Sensing*, 70(5):627–633, 2004. (Cited on pages 149 and 150.)

G. M. Foody, D. S. Boyd, and M. E. J. Cutler. Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment*, 85:463–474, 2003. (Cited on page 63.)

K. Fukunaga. *Introduction to statistical pattern recognition*. Academic press, 1990. (Cited on page 38.)

M. G. Genton. Classes of kernels for machine learning: a statistics perspective. *The Journal of Machine Learning Research*, 2:299–312, 2002. (Cited on page 38.)

T. Gevers and H. Stokman. Robust histogram construction from color invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1): 113–118, 2004. (Cited on page 25.)

A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock. Imaging spectrometry for Earth remote sensing. *Science*, 228(4704):1147–1153, 1985. (Cited on page 18.)

L. Gomez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla. Mean map kernel methods for semisupervised cloud classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48(1):207–220, 2010. (Cited on pages 51 and 68.)

B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2066–2073. IEEE, 2012. (Cited on page 56.)

R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice Hall, 2nd edition, 2002. (Cited on pages 19 and 24.)

R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 999–1006. IEEE, 2011. (Cited on page 56.)

R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. (Cited on page 56.)

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005. (Cited on page 90.)

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. (Cited on page 51.)

J. H. Ham, D. D. Lee, and L. K. Saul. Learning high dimensional correspondences from low dimensional manifolds. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003. (Cited on page 56.)

Bruce Hapke, Dominick DiMucci, Robert Nelson, and William Smythe. The cause of the hot spot in vegetation canopies and soils: Shadow-hiding versus coherent backscatter. *Remote Sensing of Environment*, 58(1):63–68, 1996. (Cited on page 112.)

R. M. Haralick, K. Shanmugam, and I. H. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973. (Cited on pages 21 and 74.)

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 2. Springer, 2009. (Cited on pages 36 and 40.)

G. E. Hilley, R. Bürgmann, A. Ferretti, F. Novali, and F. Rocca. Dynamics of slow-moving landslides from permanent scatterer analysis. *Science*, 304(5679):1952–1955, 2004. (Cited on page 12.)

C. G. Homer, R. D. Ramsey, T. C. Edwards Jr, and A. Falconer. Landscape cover-type modeling using a multi-scene thematic mapper mosaic. *Photogrammetric Engineering and Remote Sensing*, 63(1):59–67, 1997. (Cited on page 23.)

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441 and 498–520, 1933. (Cited on page 41.)

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936. (Cited on page 65.)

J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 601–608. MIT Press, Cambridge, MA, 2007. (Cited on pages 51, 55, and 68.)

X. Huang and L. Zhang. An adaptive mean-shift analysis approach for object extraction and classification from urban hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 46(12):4173–4185, 2008. (Cited on page 21.)

G. F. Hughes. On the mean accuracy of statistical pattern recognition. *IEEE Transactions on Information Theory*, IT-14(1):55–63, 1968. (Cited on page 41.)

S. Inamdar, F. Bovolo, and L. Bruzzone. Multidimensional probability density function matching for preprocessing of multitemporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1243–1252, 2008. (Cited on page 25.)

K. I. Itten, F. Dell'Endice, A. Hueni, M. Kneubühler, D. Schläpfer, D. Odermatt, F. Seidel, S. Huber, J. Schopfer, T. Kellenberger, Y. Bühler, P. D'Odorico, J. Nieke, E. Alberti, and K. Meuleman. APEX - the hyperspectral ESA airborne prism experiment. *Sensors*, 8(10): 6235–6259, 2008. (Cited on page 18.)

M. Jaboyedoff, T. Oppikofer, A. Abellán, M.-H. Derron, A. Loye, R. Metzger, and A. Pedrazzini. Use of LIDAR in landslide investigations: a review. *Natural Hazards*, 61(1):5–28, 2012. (Cited on page 12.)

Q. Jackson and D. Landgrebe. An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Transactions on Geoscience and Remote Sensing*, 39(12):2664–2679, 2001. (Cited on page 68.)

J. P. Jacobs, G. Thoonen, D. Tuia, G. Camps-Valls, B. Haest, and P. Scheunders. Domain adaptation with hidden markov random fields. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Melbourne, Australia, 2013. (Cited on page 66.)

S. Jegelka, A. Gretton, B. Schölkopf, B. K. Sriperumbudur, and U. von Luxburg. Generalized clustering via kernel embeddings. In *Annual German Conference on Artificial Intelligence*, Paderborn, Germany, 2009. (Cited on page 52.)

J. R. Jensen and D. C. Cowen. Remote sensing of urban/suburban infrastructure and socio-economic attributes. *Photogrammetric Engineering and Remote Sensing*, 65:611–622, 1999. (Cited on page 12.)

Y. Jhung and P. H. Swain. Bayesian contextual classification based on modified m-estimates and markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 34 (1):67–75, 1996. (Cited on page 21.)

Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1697–1704. IEEE, 2011. (Cited on page 44.)

P. Jonsson and L. Eklundh. Seasonality extraction by function fitting to time-series of satellite sensor data. *IEEE Transactions on Geoscience and Remote Sensing*, 40(8):1824–1832, 2002. (Cited on page 21.)

K. E. Joyce, S. E. Belliss, S. V. Samsonov, S. J. McNeill, and P. J. Glassey. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 33(2):183–207, 2009. (Cited on page 12.)

G. Jun and J. Ghosh. An efficient active learning algorithm with knowledge transfer for hyperspectral data analysis. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, pages I–52–I–55, Boston, MA, USA, 2008. (Cited on page 64.)

G. Jun and J. Ghosh. Spatially adaptive classification of land cover with remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(7):2662–2673, 2011. (Cited on page 68.)

G. Jun and J. Ghosh. Semisupervised learning of hyperspectral data with unknown land-cover classes. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):273–282, 2013. (Cited on page 68.)

M. Kanevski, G. Christakos, V. Demyanov, L. Foresti, C. Kaiser, M. Maignan, A. Pozdnoukhov, R. Purves, F. Ratle, E. Savelieva, R. Tapia, V. Timonin, and D. Tuia. *Advanced Mapping of Environmental Data*. ISTE Ltd and John Wiley & Sons, Inc., 2008. (Cited on pages 30 and 33.)

M. Kanevski, A. Pozdnoukhov, and V. Timonin. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press, 2009. (Cited on page 30.)

S. Q. Kidder and T. H. Vonder Haar. *Satellite meteorology: an introduction*, volume 466. Academic press San Diego, 1995. (Cited on page 12.)

W. Kim and M. Crawford. Adaptive classification for hyperspectral image data using manifold regularization kernel machines. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4110–4121, 2010. (Cited on page 66.)

W. Kim, M. M. Crawford, and J. Ghosh. Spatially adapted manifold learning for classification of hyperspectral imagery with insufficient labeled data. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, pages I–213, Boston, MA, USA, 2008. (Cited on page 66.)

J. Knorn, A. Rabe, V. C. Radeloff, T. Kuemmerle, J. Kozak, and P. Hostert. Land cover mapping of large areas using chain classification of neighboring landsat satellite images. *Remote Sensing of Environment*, 113(5):957–964, 2009. (Cited on pages 21 and 64.)

S. Kong and D. Wang. A brief summary of dictionary learning based approach for classification. *arXiv preprint arXiv:1205.6544*, 2012. (Cited on page 44.)

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. (Cited on page 50.)

B.-C. Kuo and D. A. Landgrebe. Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geoscience and Remote Sensing*, 42(5):1096–1105, 2004. (Cited on page 65.)

D. W. Lamb and R. B. Brown. Remote-sensing and mapping of weeds in crops. *Journal of Agricultural Engineering Research*, 78(2):117–125, 2001. (Cited on page 12.)

J. M. Leiva-Murillo, L. Gómez-Chova, and G. Camps-Valls. Multitask remote sensing data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):151–161, 2013. (Cited on page 68.)

J. Li, H. Zhang, Y. Huang, and L. Zhang. Hyperspectral image classification by nonlocal joint collaborative representation with a locally adaptive dictionary. *IEEE Transactions on Geoscience and Remote Sensing*, 52(6):3707–3719, 2014. (Cited on page 44.)

W. Li, S. Prasad, J. E. Fowler, and L. M. Bruce. Locality-preserving discriminant analysis in kernel-induced feature spaces for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 8(5):894–898, 2011. (Cited on page 65.)

G. Licciardi, F. Pacifici, D. Tuia, S. Prasad, T. West, F. Giacco, C. Thiel, J. Inglada, E. Christophe, J. Chanussot, and P. Gamba. Decision fusion for the classification of hyperspectral data: Outcome of the 2008 GRSS data fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3857–3865, 2009. (Cited on page 156.)

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. (Cited on page 50.)

N. Longbotham, C. Chaapel, L. Bleiler, C. Padwick, W. J. Emery, and F. Pacifici. Very high resolution multiangle urban classification analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 50(4):1155–1170, 2012a. (Cited on pages 22, 113, 117, and 158.)

N. Longbotham, F. Pacifici, and W. Emery. In-track multi-angle model portability of multi-spectral land-cover classification using very high spatial resolution data. In *Proceeding of the SPIE conference on Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery*, Baltimore, USA, 2012b. (Cited on page 113.)

D. Lunga, S. Prasad, M. Crawford, and O. Ersoy. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine*, 31(1):55–66, 2014. (Cited on page 65.)

T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. Active learning to recognize multiple types of plankton. *Journal of Machine Learning Research*, 6:589–613, 2005. (Cited on pages 69 and 72.)

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 2008. (Cited on page 44.)

I. Manakos and M. Braun. *Land Use and Land Cover Mapping in Europe*. Springer, 2014. (Cited on page 12.)

F. Mantovani, R. Soeters, and C. J. Van Westen. Remote sensing techniques for landslide studies and hazard zonation in europe. *Geomorphology*, 15(3):213–225, 1996. (Cited on page 12.)

D. Marcos Gonzalez, F. de Morsier, G. Matasci, D. Tuia, and J.-P. Thiran. Hierarchical sparse representation for dictionary-based classification of hyperspectral images. In *Proceedings of the IEEE Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS)*, Lausanne, Switzerland, 2014. (Cited on pages 9 and 45.)

A. Margolis. A literature review of domain adaptation with unlabeled data. Technical report, University of Washington, USA, 2011. (Cited on page 53.)

J.-F. Mas. Monitoring land-cover changes: a comparison of change detection techniques. *International Journal of Remote Sensing*, 20(1):139–152, 1999. (Cited on page 12.)

G. Matasci, D. Tuia, and M. Kanevski. Domain separation for efficient adaptive active learning. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3716–3719, Vancouver, Canada, 2011a. (Cited on page 7.)

G. Matasci, M. Volpi, D. Tuia, and M. Kanevski. Transfer Component Analysis for domain adaptation in image classification. In *Proceeding of the SPIE Remote Sensing conference on Image and Signal Processing for Remote Sensing*, Prague, Czech Republic, 2011b. (Cited on page 8.)

G. Matasci, D. Tuia, and M. Kanevski. SVM-based boosting of active learning strategies for efficient domain adaptation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(5):1335–1343, 2012. (Cited on page 7.)

G. Matasci, L. Bruzzone, M. Volpi, D. Tuia, and M. Kanevski. Investigating feature extraction for domain adaptation in remote sensing image classification. In *Proceedings of the International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Barcelona, Spain, 2013a. (Cited on page 8.)

G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia. Statistical assessment of dataset shift and model portability in multi-angle in-track image acquisitions. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4134–4137, Melbourne, Australia, 2013b. (Cited on page 8.)

G. Matasci, F. de Morsier, M. Kanevski, and D. Tuia. Domain adaptation in remote sensing through cross-image synthesis with dictionaries. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Québec City, Canada, 2014. (Cited on page 9.)

G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia. Semisupervised Transfer Component Analysis for domain adaptation in remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, Accepted. (Cited on page 7.)

G. Matasci, N. Longbotham, F. Pacifici, M. Kanevski, and D. Tuia. Understanding angular effects in VHR in-track multi-angle image sequences and their consequences on urban land-cover model portability. *ISPRS Journal of Photogrammetry and Remote Sensing*, Submitted. (Cited on page 8.)

P. Mather and B. Tso. *Classification methods for remotely sensed data*. CRC press, 2003. (Cited on page 20.)

K. McLaren. The development of the CIE 1976 (l* a* b*) uniform colour space and colour-difference formula. *Journal of the Society of Dyers and Colourists*, 92(9):338–341, 1976. (Cited on page 25.)

F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8): 1778–1790, 2004. (Cited on page 20.)

M. S. Moran, Y. Inoue, and E. M. Barnes. Opportunities and limitations for image-based remote sensing in precision crop management. *Remote sensing of Environment*, 61(3): 319–346, 1997. (Cited on page 12.)

J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012. (Cited on page 48.)

G. Moser, S. B. Serpico, and J. A. Benediktsson. Land-cover mapping by markov modeling of spatial–contextual information in very-high-resolution remote sensing images. *Proceedings of the IEEE*, 101(3):631–651, 2013. (Cited on page 21.)

A. A. Mueller, A. Hausold, and P. Strobl. HySens-DAIS/ROSIS imaging spectrometers at DLR. In *Proceeding of the SPIE Remote Sensing conference on Image and Signal Processing for Remote Sensing*, pages 225–235, 2002. (Cited on page 18.)

H. Nagendra. Using remote sensing to assess biodiversity. *International Journal of Remote Sensing*, 22(12):2377–2400, 2001. (Cited on page 12.)

P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 100(9):917–922, 1977. (Cited on page 65.)

G. H. Nguyen, S. L. Phung, and A. Bouzerdoum. Efficient SVM training with reduced weighted samples. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–5, San Jose, CA, USA, 2010. (Cited on page 70.)

J. Ni, Q. Qiu, and R. Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 692–699, 2013. (Cited on page 57.)

A. A. Nielsen. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–78, 2007. (Cited on page 65.)

A. A. Nielsen and M. J. Canty. Kernel principal component analysis for change detection. In *Proceeding of the SPIE Remote Sensing conference*, Cardiff, 2008. (Cited on pages 42 and 66.)

A. A. Nielsen and J. S. Vestergaard. A kernel version of multivariate alteration detection. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3451–3454, 2013. (Cited on page 65.)

A. A. Nielsen, K. Conradsen, and J. J. Simpson. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64:1–19, 1998. (Cited on pages 24 and 65.)

I. Olthof, C. Butson, and R. Fraser. Signature extension through space for northern land-cover classification: A comparison of radiometric correction methods. *Remote Sensing of Environment*, 95(3):290–302, 2005. (Cited on pages 21 and 63.)

F. Pacifici. An automatic atmospheric compensation algorithm for very high spatial resolution imagery and its comparison to quac and flaash. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Melbourne, Australia, 2013. (Cited on page 114.)

F. Pacifici, N. Longbotham, and W. J. Emery. The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, 2014. (Cited on page 24.)

M. Pal. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222, 2005. (Cited on page 20.)

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. (Cited on pages 21, 47, 48, and 53.)

S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Chicago, IL, 2008. (Cited on page 55.)

S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the International Conference on World Wide Web*, pages 751–760, 2010. (Cited on page 56.)

S. J. Pan, I. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. (Cited on pages 56, 86, 88, 89, and 93.)

B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics, 2002. (Cited on page 56.)

M. Pax-Lenney, C. E. Woodcock, S. A. Macomber, S. Gopal, and C. Song. Forest mapping with a generalized classifier and Landsat TM data. *Remote Sensing of Environment*, 77 (3):241–250, 2001. (Cited on page 63.)

C. Persello and L. Bruzzone. Active learning for domain adaptation in the supervised classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4468–4483, 2012. (Cited on page 64.)

M. Pesaresi and J. A. Benediktsson. A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(2):309–320, 2001. (Cited on pages 21 and 75.)

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press, 1999. (Cited on page 72.)

A. Plaza, J. A. Benediktsson, J. W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J. C. Tilton, and G. Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009. (Cited on page 18.)

D. M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011. (Cited on page 150.)

E. Puttonen, J. Suomalainen, T. Hakala, and J. Peltoniemi. Measurement of reflectance properties of asphalt surfaces and their usability as reference targets for aerial photos. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2330–2339, 2009. (Cited on page 113.)

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009. (Cited on page 48.)

B. Rabus, M. Eineder, A. Roth, and R. Bamler. The shuttle radar topography mission – a new class of digital elevation models acquired by spaceborne radar. *Journal of Photogrammetry and Remote Sensing*, 57(4):241–262, 2003. (Cited on page 12.)

H. Rahman, B. Pinty, and M. M. Verstraete. Coupled surface-atmosphere reflectance (CSAR) model: 2. Semiempirical surface model usable with NOAA advanced very high resolution radiometer data. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 98(D11): 20791–20801, 1993. (Cited on page 113.)

P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, Los Angeles, CA, USA, 2010. (Cited on page 55.)

S. Rajan, J. Ghosh, and M. M. Crawford. Exploiting class hierarchies for knowledge transfer in hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 44(11): 3408–3417, 2006. (Cited on page 67.)

S. Rajan, J. Ghosh, and M. M. Crawford. An active learning approach to hyperspectral data classification. *IEEE Transactions on Geoscience and Remote Sensing*, 46(4):1231–1242, 2008. (Cited on pages 64 and 154.)

B. Raup, A. Kääb, J. S. Kargel, M. P. Bishop, G. Hamilton, E. Lee, F. Paul, F. Rau, D. Soltesz, S. J. S. Khalsa, M. Beedle, and Helm C. Remote sensing and GIS technology in the global land ice measurements from space (GLIMS) project. *Computers and Geosciences*, 33(1):104–125, 2007. (Cited on page 12.)

B. C. Reed, J. F. Brown, D. VanderZee, T. R. Loveland, J. W. Merchant, and D. O. Ohlen. Measuring phenological variability from satellite imagery. *Journal of Vegetation Science*, 5(5):703–714, October 1994. (Cited on page 22.)

J. A. Richards and X. Jia. *Remote sensing digital image analysis*, volume 3. Springer, 1999. (Cited on pages 11 and 17.)

J.-L. Roujean, M. Leroy, and P.-Y. Deschamps. A bidirectional reflectance model of the Earth's surface for the correction of remote sensing data. *Journal of Geophysical Research, D: Atmospheres (1984–2012)*, 97(D18):20455–20468, 1992. (Cited on page 112.)

F. F. Sabins. Remote sensing for mineral exploration. *Ore Geology Reviews*, 14(3):157–183, 1999. (Cited on page 12.)

A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, and S. L. DuVall. Active supervised domain adaptation. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 97–112, Athens, Greece, 2011. (Cited on page 55.)

S. Satpal and S. Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 224–235, Warsaw, Poland, 2007. (Cited on page 56.)

G. Schaepman-Strub, M. E. Schaepman, T. H. Painter, S. Dangel, and J.V. Martonchik. Reflectance quantities in optical remote sensing–definitions and case studies. *Remote Sensing of Environment*, 103(1):27–42, 2006. (Cited on pages 15 and 112.)

R. E. Schapire. A brief introduction to boosting. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1401–1406, 1999. (Cited on page 54.)

B. Schölkopf and A. J. Smola. *Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002. (Cited on page 36.)

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. (Cited on page 42.)

J. R. Schott. *Remote sensing: the image chain approach*. Oxford: Oxford University Press, 2007. (Cited on pages 14 and 112.)

J. R. Schott, C. Salvaggio, and W. J. Volchok. Radiometric scene normalization using pseudoinvariant features. *Remote Sensing of Environment*, 26(1):1–16, October 1988. (Cited on page 24.)

R. A. Schowengerdt. *Remote Sensing: Models and Methods for Image Processing*. Academic Press, Inc., 3rd edition, 2007. (Cited on pages 11, 12, 13, 17, 22, 23, and 111.)

S. B. Serpico and L. Bruzzone. A new search algorithm for feature selection in hyperspectral remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7): 1360–1367, 2001. (Cited on page 50.)

B. M. Shahshahani and D. A. Landgrebe. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, 32(5):1087–1095, 1994. (Cited on page 66.)

J. Shan and C. K. Toth. *Topographic laser ranging and scanning: principles and processing*. CRC Press, 2008. (Cited on page 12.)

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. (Cited on pages 34 and 38.)

S. Shekhar, V. M. Patel, H. V. Nguyen, and R. Chellappa. Generalized domain-adaptive dictionaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 361–368, 2013. (Cited on page 57.)

X. Shi, W. Fan, and J. Ren. Actively transfer domain knowledge. In *Proceedings of the European Conference on Machine Learning (ECML)*, pages 342–357, Berlin, Germany, 2008. (Cited on page 55.)

C. Simmer and S. A. W. Gerstl. Remote sensing of angular characteristics of canopy reflectances. *IEEE Transactions on Geoscience and Remote Sensing*, GE-23(5):648–658, 1985. (Cited on pages 14 and 112.)

C. Song, C. E. Woodcock, K. C. Seto, M. Pax-Lenney, and S. A. Macomber. Classification and change detection using Landsat TM data: When and how to correct atmospheric effects? *Remote Sensing of Environment*, 75(2):230–244, 2001. (Cited on page 24.)

H. Song, B. Huang, K. Zhang, and H. Zhang. Spatio-spectral fusion of satellite images based on dictionary-pair learning. *Information Fusion*, 18:148–160, 2014. (Cited on page 130.)

A. H. Strahler. The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment*, 10(2):135–163, 1980. (Cited on page 20.)

M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems (NIPS)*, volume 7, pages 1433–1440, 2007. (Cited on page 55.)

Z. Sun, C. Wang, H. Wang, and J. Li. Learn multiple-kernel SVMs for domain adaptation in hyperspectral data. *IEEE Geoscience and Remote Sensing Letters*, 2013. (Cited on page 68.)

P. H. Swain. Pattern recognition: a basis for remote sensing data analysis. Technical report, LARS Information Note 111572, LARS, Purdue University, West Lafayette, Indiana., 1972. (Cited on page 30.)

C. Tao, Y. Tang, C. Fan, and Z. Zou. Hyperspectral imagery classification based on rotation-invariant spectral-spatial feature. *IEEE Geoscience and Remote Sensing Letters*, 11(5): 980–984, 2014. (Cited on page 133.)

P. M. Teillet, B. Guindon, and D. G. Goodenough. On the slope-aspect correction of multispectral scanner data. *Canadian Journal of Remote Sensing*, 8(2):84–106, 1982. (Cited on page 22.)

S. Thrun and L. Pratt. Learning to learn: introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998. (Cited on page 48.)

I. Tosic and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, 2011. (Cited on page 43.)

G. T. Toussaint. Some inequalities between distance measures for feature evaluation. *IEEE Transactions on Computers*, 21(4):409–410, 1972. (Cited on page 50.)

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007. (Cited on page 43.)

D. Tuia and G. Camps-Valls. Semisupervised remote sensing image classification with cluster kernels. *IEEE Geoscience and Remote Sensing Letters*, 6(2):224–228, 2009. (Cited on page 20.)

D. Tuia, F. Pacifici, M. Kanevski, and W. J. Emery. Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3866–3879, 2009a. (Cited on page 21.)

D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47

(7):2218–2232, 2009b. (Cited on page 64.)

D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski. Learning relevant image features with multiple-kernel classification. *IEEE Transactions on Geoscience and Remote Sensing*, 48 (10):3780–3791, 2010. ISSN 0196-2892. doi: 10.1109/TGRS.2010.2049496. (Cited on page 65.)

D. Tuia, E. Pasolli, and W. J. Emery. Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment*, 115(9):2232–2242, 2011a. (Cited on page 64.)

D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí. A survey of active learning algorithms for remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011b. (Cited on pages 21, 64, and 70.)

D. Tuia, J. Muñoz-Marí, L. Gomez-Chova, and J. Malo. Graph matching for adaptation in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):329–341, 2013a. (Cited on page 66.)

D. Tuia, M. Trolliet, and M. Volpi. Multisensor alignment of image manifolds. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Melbourne, Australia, 2013b. (Cited on page 66.)

D. Tuia, E. Merenyi, X. Jia, and M. Grana-Romay. Foreword to the special issue on machine learning for remote sensing data processing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4):1007–1011, 2014. (Cited on page 30.)

D. Tuia, M. Volpi, M. Trolliet, and G. Camps-Valls. Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, In press. (Cited on page 66.)

G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter. The airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, 44(2):127–143, 1993. (Cited on page 18.)

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998. (Cited on pages 31, 32, 34, and 37.)

J. Verrelst, M. E. Schaepman, and J. G. P. W. Clevers. Fusing Minnaert-k parameter with spectral unmixing for forest heterogeneity mapping using CHRIS-PROBA data. In *Proceedings of the IEEE Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS)*, pages 1–4, 2009. (Cited on page 113.)

J. Verrelst, M. E. Schaepman, and J. G. P. W. Clevers. Spectrodirectional Minnaert-k retrieval using CHRIS-PROBA data. *Canadian Journal of Remote Sensing*, 36(6):631–644, 2010. (Cited on page 113.)

M. Volpi, G. Matasci, D. Tuia, and M. Kanevski. Enhanced change detection using nonlinear feature extraction. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 6757–6760, Munich, Germany, 2012a. (Cited on pages 8 and 66.)

M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection with kernels. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1026–1030, 2012b. (Cited on page 20.)

M. Volpi, D. Tuia, and M. Kanevski. Memory-based cluster sampling for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3096–3106, 2012c. (Cited on page 64.)

M. Volpi, F. de Morsier, G. Camps-Valls, M. Kanevski, and D. Tuia. Multi-sensor change detection based on nonlinear canonical correlations. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2013. (Cited on page 65.)

M. Volpi, G. Matasci, M. Kanevski, and D. Tuia. Semi-supervised multiview embedding for hyperspectral data classification. *Neurocomputing*, In press. (Cited on page 8.)

C. Wang and S. Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the International Conference on Machine learning (ICML)*, pages 1120–1127, 2008. (Cited on page 56.)

C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, page 3, 2009.

(Cited on page 56.)

C. Wang and S. Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Barcelona, 2011. (Cited on pages 56, 66, and 130.)

S. Wang, D. B. Elliott, J. B. Campbell, R. W. Erich, and R. M. Haralick. Spatial reasoning in remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, (1): 94–101, 1983. (Cited on page 21.)

S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2216–2223, 2012. (Cited on pages 57, 130, 131, and 134.)

Z. Wang, N. M. Nasrabadi, and T.S. Huang. Spatial-spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8):4808–4822, 2014. (Cited on page 44.)

A. Wehr and U. Lohr. Airborne laser scanning – an introduction and overview. *Journal of Photogrammetry and Remote Sensing*, 54(2):68–82, 1999. (Cited on page 12.)

G. J. Well, R. J. Graf, and L. M. Forister. Investigations of hazardous waste sites using thermal ir and ground penetrating radar. *Photogrammetric Engineering and Remote Sensing*, 60 (8):999–1005, 1994. (Cited on page 12.)

Q. Weng. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment*, 117:34–49, 2012. (Cited on page 12.)

C. E. Woodcock, S. A. Macomber, M. Pax-Lenney, and W. B. Cohen. Monitoring large areas for forest change using Landsat: Generalization across space, time and Landsat sensors. *Remote Sensing of Environment*, 78(1-2):194–203, 2001. (Cited on pages 21 and 63.)

J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010. (Cited on page 43.)

H. L. Yang and M. M. Crawford. Manifold alignment for multitemporal hyperspectral image classification. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4332–4335, 2011. (Cited on page 66.)

H. L. Yang and M. M. Crawford. Exploiting spectral-spatial proximity for classification of hyperspectral data on manifolds. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4174–4177, 2012. (Cited on page 65.)

J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the International Conference on Multimedia*, pages 188–197, 2007a. (Cited on page 57.)

J. Yang, R. Yan, and A. G. Hauptmann. Adapting SVM classifiers to data with shifted distributions. In *Proceedings of the IEEE International Conference on Data Mining Workshops*, pages 69–76, Washington, DC, USA, 2007b. (Cited on page 57.)

J. Yang, P. Gong, R. Fu, M. Zhang, J. Chen, S. Liang, B. Xu, J. Shi, and R. Dickinson. The role of satellite remote sensing in climate change studies. *Nature climate change*, 3(10): 875–883, 2013. (Cited on page 12.)

M. Yang, L. Zhang, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1601–1604, 2010. (Cited on page 44.)

Z. Yang and R. Mueller. Heterogeneously sensed imagery radiometric response normalization for citrus grove change detection. In *Proceeding of the SPIE Optics East conference*, volume 6761, Boston, MA, USA, 2007. (Cited on page 129.)

Y. Zheng, W. Dong, and E. P. Blasch. Qualitative and quantitative comparisons of multispectral night vision colorization techniques. *Optical Engineering*, 51(8):087004–1, 2012. (Cited on page 25.)

This Thesis was typeset with LaTeX using the `classicthesis` package available at http://code.google.com/p/classicthesis/. Thank you André Miede, a nice postcard of Lausanne is coming your way!